DOCUMENT RESUME

ED 441 652 RC 022 459

AUTHOR	Figueroa, Richard A.; Hernandez, Sonia
TITLE	Testing Hispanic Students in the United States: Technical and Policy Issues.
INSTITUTION	President's Advisory Commission on Educational Excellence for Hispanic Americans (ED), Washington, DC.
PUB DATE	2000-05-00
NOTE	71p.
AVAILABLE FROM	ED Pubs, P.O. Box 1398, Jessup, MD 20794-1398. Tel: 877-433-7827 (Toll Free). For full text: http://www.ed.gov/offices/OIIA/Hispanic/rr/.
PUB TYPE	Information Analyses (070) -- Reports - Evaluative (142)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Academic Standards; Accountability; Achievement Tests; *Bilingual Students; Educational Assessment; Educational Diagnosis; *Educational Testing; Elementary Secondary Education; *Hispanic American Students; Limited English Speaking; Mexican American Education; Special Education; Student Evaluation; *Test Bias; *Testing Problems

ABSTRACT

In the rush to implement world-class standards supported by systems of accountability in public schools, education leaders have compromised the educational future of Hispanic students by making high-stakes decisions based on inaccurate and inadequate testing information. State policies often require that Hispanic students be assessed in English with tests they don't understand or in Spanish with alternative but less rigorous tests, whether or not they are receiving instruction in Spanish. Neither approach produces accurate information about student learning, but the results are often used to hold students, rather than educators or the public school systems accountable. Where exemptions from testing exist, Hispanics disappear from accountability systems. This report aims to bring attention to the growing crisis of "invisible" Hispanic students and provide guidance on rectifying the conditions that allow Hispanic students to be wrongly measured and unaccounted for in their own schools. Chapters cover: (1) basic questions about cultural and linguistic influences on test scores; (2) historical context of educational and psychometric testing of linguistic and cultural minority groups; (3) standards for educational and psychological testing and Office for Civil Rights guidelines; (4) six sources of test bias with Hispanic test-takers; (5) Hispanic performance on achievement tests and efforts by the National Research Council to address the testing of English language learners; (6) diagnostic testing for special education; (7) other types of testing; and (8) a summary of 7 options for the testing of Hispanic children, followed by 18 recommendations to the Office for Civil Rights, educators, and researchers. An appendix contains standards for testing individuals of diverse linguistic backgrounds. (Contains approximately 180 references.) (SV)

Reproductions supplied by EDRS are the best that can be made
from the original document.

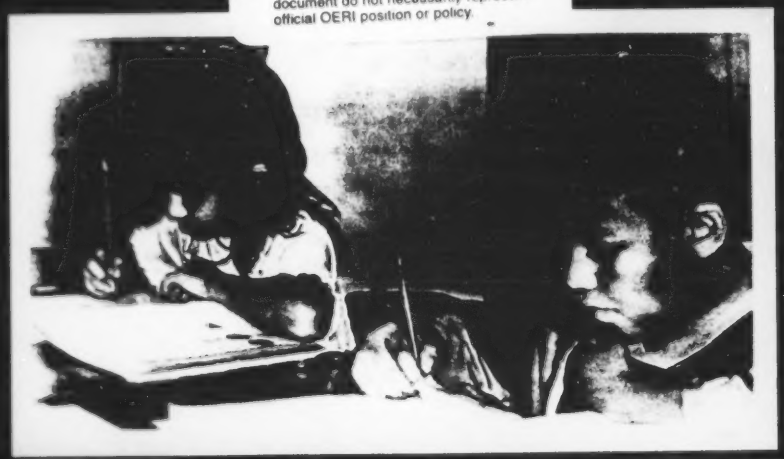
RC

President's Advisory Commission on Educational Excellence for Hispanic Americans

ED 441 652

Testing Hispanic Students in the United States: Technical and Policy Issues

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.



RC022459

BEST COPY AVAILABLE

PRESIDENT'S ADVISORY COMMISSION ON EDUCATIONAL EXCELLENCE FOR HISPANIC AMERICANS

Guillermo Linares (Chair)*
New York, New York

Sonia Hernández (Vice Chair)
Sacramento, CA

Erlinda Paiz Archuleta*
Denver, Colorado

Cecilia Preciado Burciaga
Seaside, California

George Castro*
San Jose, California

Darlene Chavira Chávez
Tuscon, Arizona

David J. Cortiella
Boston, Massachusetts

Miriam Cruz
Washington, D.C.

Juliet Villareal García
Brownsville, Texas

José González
San Juan, Puerto Rico

Cipriano Muñoz*
San Antonio, Texas

Harry P. Pachón
Claremont, California

Eduardo J. Padrón
Miami, Florida

Janice Petrovich
New York, New York

Gloria Rodríguez
San Antonio, Texas

Waldemar Rojas*
Dallas, Texas

Isaura Santiago-Santiago
New York, New York

John Phillip Santos
New York, New York

Samuel Vigil*
Las Vegas, New Mexico

Diana Wasserman
Ft. Lauderdale, Florida

Rubén Zacarías*
Los Angeles, California

(*Members, Commission Assessment Committee)

White House Initiative on Educational Excellence for Hispanic Americans Staff

Sarita E. Brown
Executive Director

Deborah A. Santiago
Deputy Director

Richard Toscano
Special Assistant for Interagency Affairs

Julie Laurel
Policy Analyst

Deborah M. Montoya
Assistant to the Executive Director

Danielle Gonzales
Policy Intern

TESTING HISPANIC STUDENTS IN THE UNITED STATES: TECHNICAL AND POLICY ISSUES

President's Advisory Commission
on Educational Excellence
for Hispanic Americans

May 2000

Prepared by: Richard A. Figueroa, University of California at Davis
Sonia Hernandez, California Department of Education

ACKNOWLEDGEMENTS

A very special thanks is extended to Dianna Gutiérrez. We are indebted for the tremendous amount of research that she directed in the preparation of this report. Commendations are extended to her assistant, Marysol Flores.

We must also acknowledge the help of many colleagues throughout the United States. They were all generous, prompt and supportive in their efforts.

The Professional Staff of WestEd: Stanley Rabinowitz, Rose Marie Fontana, Sri Ananda, Robert Linquanti

Guadalupe Valdés, Stanford University
Janette Klinger, University of Miami
José Cintrón, California State University Sacramento
Lila Jacobs, California State University Sacramento
Linda Murai, Sacramento County Office of Education
Pedro Pedraza, Center for Puerto Rican Studies, Hunter College
Diane August, August & Associates
Jon Sandoval, University of California at Davis
Eugene García, University of California at Berkeley
Alfredo Artiles, University of California at Los Angeles
Maria Trejo and David Dolson, California Department of Education

CONTENTS

Foreword iv

Chapter 1: Introduction 1

Chapter 2 : The Historical Context 3

Chapter 3: Testing Standards and Office for Civil Rights (OCR) Guidelines 7

Chapter 4: Bias 19

Chapter 5: Educational Accountability 27

Chapter 6: Diagnostic Testing for Special Education 35

Chapter 7: Other Types of Testing 39

Chapter 8: Summary and Recommendations 43

References 49

Appendix A 61

FOREWORD

There is no more promising reform in public education today than the standards-based movement. It is not only the most widely accepted school change process, it also offers the greatest probability for leveling the playing field for all children, by clearly stating expectations for instruction, assessing the progress of each child toward achieving the standards, and holding schools accountable for student learning. Where these three core elements of a standards-based system--clear expectations, assessment and accountability--are in place, students experience success as never before. This is especially true for the growing Hispanic student population in the United States, which has traditionally had limited access to rigorous mainstream instruction.

But in the current rush to implement world-class standards supported by systems of accountability in the nation's public schools, state education leaders have compromised the educational future of Hispanic students by making high-stakes decisions based on inaccurate and inadequate testing information. Hundreds of thousands of Hispanic students, many lacking functional fluency in English, are assessed with a myriad of tests entirely in English and, oftentimes, only in English. The resulting data is used to determine high-stakes decisions, such as for student promotion or retention, or high school graduation--but rarely for the purposes of true accountability. When it comes to holding schools accountable for the academic achievement of our students, states allow Hispanic youngsters to become invisible inside the very system charged with educating them.

State policies often require that Hispanic students be assessed in English with tests they may not even understand or with alternative but less rigorous tests in Spanish whether or not they are receiving instruction in that language. While neither approach produces accurate information about student learning, the resulting data is often used to hold students accountable for their own success, rather than the educators or the public school systems.

Who should be responsible for what Hispanic students learn in school? The answer is simple: students, educators, and parents all must share the responsibility.

But what kinds of assessments should be used to provide accurate information about what students have been taught? Regrettably, the answer to this question is not as simple. It is explored in this document.

With few exceptions, students bear the weight of academic success or failure on the basis of one or two test scores. Where exemptions from testing exist, Hispanics disappear from the accountability reports, triggering both positive and negative consequences for the responsible adults in the system. Thus more than two million Hispanic students in the United States are underrepresented or absent from the rolls of students who are counted via assessment and who, therefore, count.

It is our belief that Hispanic students, whether they are English dominant or English Language Learners, should be tested with appropriate test instruments in order to be

included at all times in the states' accountability systems. If this does not occur, Hispanic children will not benefit from the powerful and promising standards movement. As the United States enters the new millennium, deliberate action by policymakers at every level must be taken to include the country's fastest growing and soon-to-be largest minority, within the bounds of systems accountability using accurate information for decisionmaking.

The purpose of this report is twofold: (1) to bring attention to the growing crisis of the "invisible" Hispanic students in public education to the nation's leaders and (2) to provide guidance to the nation and the states on taking the necessary steps to rectify the conditions that allow Hispanic students to be wrongly measured and unaccounted for in their own schools. It is our intent to help education leaders in this country choose wisely for the sake of the children.

Commission Assessment Committee--President's Advisory Commission on Educational Excellence For Hispanic Americans

Washington, D.C.
September 15, 1999

CHAPTER 1: INTRODUCTION

All forms of human mental measurement are fragile and problematic (Gould, 1981). At their best, for example, psychometric tests account for a modest 25-35 percent of the variance of what they predict (Neiser, Bodoo, Bouchard, Boykin, Brody, Ceci, Helpern, Loehlin, Perloff, Sternberg, & Urbina, 1996; Cleary, Humphreys, Kendrick, & Wesman, 1975). This is a technical ceiling that test-makers have not succeeded in breaking for nearly a century. A fundamental assumption of all testing is that the normative framework (psychometric, criterion, or rubrics-based) on which the test scores are based assumes a high degree of experiential homogeneity, cultural/linguistic similarity and equity in learning opportunities among test takers (Colvin, 1921; Woodrow, 1921; Heller, Holtzman, & Messick, 1982). Under these conditions, a test score becomes a measure that belongs pre-eminently to the individual and his or her talents, achievements, traits and predispositions. In a real sense, tests work best in a perfect democracy of monolingual and monocultural citizens.

Hispanic Americans present a massive challenge to the assumptions of tests. The vast majority has varying levels of exposure to and proficiency in Spanish, though many also come from other linguistic backgrounds (e.g., Portuguese, Catalán, Basque). Their cultural ancestries include Mexico, Latin America, Puerto Rico, Cuba, the Caribbean, Spain and Portugal. Their cultural experiences in the United States are multigenerational and reflect a broad range of acculturation levels, socioeconomic differences, and political power. So vast is their heterogeneity, that the assumptions of tests about homogeneity may well be untenable. Yet, Hispanic students and Hispanic citizens are tested every day and are compared to middle class America in the unique reification of democracy and assimilation that tests impose. But the history of testing Hispanics in the United States has never been typified by equanimity.

There are few issues in American psychology or education that are as complex or as misunderstood as the testing of Hispanic students. Two fundamental questions have challenged and continue to perplex test-makers, test-givers and test-users: Does Spanish in the home or as the primary language affect test scores? and, Do any aspects of Hispanic culture in the United States attenuate or change test outcomes?

In the 1930s, the great Mexican American psychologist, George Sánchez, addressed both issues.

> The relative responsibility of the school and of the child in the achievement of desirable goals must be examined. Is the fact that a child makes an inferior score on an intelligence test prima facie evidence that he is dull? Or is it the function of the test to reflect the inferior or different training and development with which the child was furnished by his home, his language, the culture of his people, and by his school? When the child fails in promotion is it *his* failure or has the school failed to use the proper whetstone in bringing out the true temper and quality of his steel?
>
> The school has the responsibility of supplying those experiences to the child which will make the experiences sampled by standard measures as common to him as they were to those on whom the norms of the measures were based. When the school has met the language, cultural, disciplinary, and informational lacks of the child and the child has reached a saturation point of his capacity in the assimilation of fundamental experiences

> and activities - then failure on his part to respond to tests of such experiences and activities may be considered his failure. As long as the tests do not at least sample in equal degree a state of saturation that is equal for the "norm children" and the particular bilingual child it cannot be assumed that the test is a valid one for the child. (Sánchez, 1934, pgs. 770-771)

Presently, the impact of **cultural** differences on test scores remains understudied. Most of what is known about cultural effects comes from the use of U.S.-made tests on foreign populations. Anthropologists were early consumers who believed that the scientific nature of tests made them appropriate for universal use. Very little is known about cross-cultural differences in testing, in fact, precisely because monocultural tests, when translated into the local language, yielded predominantly lower scores and Anglocentric interpretations. There are, however, some notable exceptions.

Holtzman, Díaz-Guerrero, & Schwarts (1975) conducted a longitudinal study comparing approximately 400 middle class Mexican children with 400 middle class white children from northern Texas. One of the unique aspects of this investigation was that a comprehensive attempt was made to make all the sociological, psychological and educational tests and scoring protocols appropriate for Mexican students and their families. The result was a compelling description of cultural differences as well as of the production of knowledge about how psychometric tests need to undergo a radical overhaul for crosscultural use and how cultural bias can subtly affect scores. Regrettably, this type of investigation has never been replicated with Hispanic children and their families living in the United States.

By and large, the study of cultural differences in testing has always operated from a "black box" design. Culture has resided in the "Puerto Rican", "Mexican," or "Cuban American" samples used (Valdés & Figueroa, 1994). The only cultural effect on U.S.-normed, English-language tests has been lower scores. Interestingly, these types of "black box" studies have seldom found evidence of lower test reliabilities or validities because of cultural differences (Cleary, Humphreys, Kendrick, & Wesman, 1975; Geisinger, 1992; Sandoval, Frisby, Geisinger, Scheuneman, & Grenier, 1998).

The same has not been true for the one cultural variable that has left its mark on virtually every investigation using tests with Hispanic populations. Linguistic exposure to Spanish has affected every type of psychometric test and test score given in the United States (Valdés & Figueroa, 1994). It is the one variable for which there is evidence of psychometric bias (Figueroa & García, 1995). It is the one variable that finally has drawn the attention of the scientific community as a complex disrupter of established testing policies and practices (Pellegrino, Jones, & Mitchell, 1999).

CHAPTER 2 : THE HISTORICAL CONTEXT

In 1922, Charles Brigham published *The Study of American Intelligence*, an analysis of government data from testing conducted on World War I recruits. A subset of the large sample included 11,300 foreign-born recruits who were tested with the Army Alpha (a verbal test of intelligence) and the Army Beta (a nonverbal test of intelligence). Approximately 32 percent of them had been in the United States 0-5 years, 38 percent 6-10 years, 17 percent 11-15 years, 7 percent 16-20 years, and 6 percent over 20 years. Because of its high verbal content, the Army Alpha was a good measure of English language proficiency. In spite of Brigham's tortured defense of the integrity of this sample of foreign-born recruits, there is a strong indication that after 10 years the sample reflected a different type of immigrants—those who did not return to their native countries and who in all likelihood had adapted culturally and linguistically. Taking this into account, the increase in Army Alpha scores for the first two, five-year groups was a meager .11 of a point. This is one of the first major empirical findings that showed proficiency in a language besides English systematically produces lower verbal scores--possibly for a very, very long time.

The 1920s and 1930s produced a great amount of research on ethnic minorities in the United States. Much of it came under the title of "Race Psychology" and reflected a naive use of test scores to support genetic arguments about lower intellectual potential in non-Nordic groups. A considerable amount of this published work included Hispanic test subjects whose linguistic backgrounds can generally be described as "bilingual." That is, they came from homes where Spanish was spoken and with varying and unknown degrees of proficiency in Spanish. Most of these studies were conducted on Mexican American children.

Several conclusions can be extracted from this early research on bilingual test-takers. First, the test results of bilingual individuals compared to those of monolinguals, for all age groups, consistently produced a profile of lower (English) test scores regardless of the test being used. This was most pronounced in tests of verbal intelligence, although a similar profile appeared in tests of academic achievement (Brown, 1922; Cebollero, 1936; Johnson, 1938; Koch & Simmons, 1926; Manuel, 1935; Pratt, 1929). There, English-dependent skills such as vocabulary, comprehension, sentence completion skills, analogies, essay composition, and the like were markedly low in bilingual test-takers in comparison to their arithmetic and memory skills. The effect of differences in exposure to English appeared to be unerasable (Saer, 1923), or as the Brigham study showed, virtually unerasable. This phenomenon became widely known as the "language handicap" of all immigrant test-takers. In many research publications, this provided a rationale for denigrating or eradicating bilingualism and instruction in the primary language.

Second, the psychometric properties of tests showed a curious profile. Bilingualism had no effect on the internal consistency and stability of tests, particularly indices of reliability (Figueroa, 1990). But on the critical external indices of validity, particularly predictive validity, bilingualism appeared to attenuate the power of tests

(Altus, 1945; Davenport, 1932; Feingold, 1924; Garth, 1928; Paschal & Sullivan, 1925; Pintner & Keller, 1922; Wheeler, 1932; Wood, 1929; Yoder, 1928).

Third, some anomalous data appeared. Bilingual individuals from middle or upper-middle class homes occasionally either outperformed monolingual, English speakers or did as well in test scores (Darcy, 1946; Feingold, 1924; Manuel, 1935; Pintner & Arsenian 1937). The "bilingual handicap" in effect was cured by advantaged or enriched environments and backgrounds. Clearly, however, in the early part of the 20th century, foreign-born individuals with such cultural capital were relatively rare. Also, individuals with two under-developed languages did worse on tests than individuals with a single, educationally developed foreign language (Altus, 1949; Arsenian, 1945; Smith, 1949, 1957). Another finding that to this day remains present and unexplained is the ability of bilingual individuals to do better than English speakers on recalling digits backward (a staple of IQ tests since their inception) (Darsie, 1926; Hung-Hsia, 1929; Jensen & Inouye, 1980 Luh & Wy, 1931; Manuel, 1935). Finally, on school grades the "bilingual handicap" did not materialize to the same degree or persistence as on tests (Bell, 1935; Smith, 1942).

Fourth, the psychometric, scientific community began the unfortunate procedural tradition of dealing with cultural groups as monolithic entities (e.g., Garth, 1920). The "Mexican" sample operationalized "Mexican culture". Socioeconomic and other intervening variables were often ignored. English language proficiency in test subjects remained as an uncontrolled source of error. Background factors such as educational backgrounds or the segregated nature of public schooling for bilingual students were overlooked. The methodological flaws in the design of studies with bilingual persons, in effect, were substantial and virtually precluded reasonable inferences or attributions. Many of these design flaws continue (e.g., MacMillan, Gresham, & Bocian, 1998; Sandoval, 1979).

For test users, however, the "language handicap" produced several innovations or, in the current lexicon, a series of accommodations. The testing community came to believe that nonverbal tests of mental ability were free of linguistic factors and were culturally neutral (Brigham, 1922). To this day, they are seen as culture fair measures of intelligence, mental aptitudes or personality. Tests were often simply translated without conducting norming studies (Lester, 1929; Mitchell, 1937; Paschal & Sullivan, 1925). These translations were used for research purposes and for conducting actual assessments. Ethnic norms were occasionally produced for some bilingual groups (Ammons & Aguero, 1950; Luh & Wy, 1931). Many caveats and precautions on the use of tests with bilingual subjects were voiced. For example, Charles Brigham, the father of the modern SAT (Scholastic Aptitude Test) and the principal investigator in "The Study of American Intelligence," also concluded that:

> For purposes of comparing individuals or groups, it is apparent that tests in the vernacular [English] must be used only with individuals having equal opportunity to acquire the vernacular of the test. This requirement precludes the use of such tests in making comparative studies of individuals brought up in homes in which the vernacular of the test is not used, or in which two vernaculars are used. The last condition is frequently violated

> here in studies of children born in this country whose parents speak another tongue. It is important as the effects of bilingualism are not entirely known. (Brigham, 1930, pg. 165)

Some 70 years later, "the effects of bilingualism [still] are not entirely known." What has changed is that there are more caveats about testing bilinguals.

CHAPTER 3: TESTING STANDARDS AND OFFICE FOR CIVIL RIGHTS (OCR) GUIDELINES

The most important and potentially powerful set of regulations and policies on the development and use of tests in the United States is the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999). They are the "standard of the industry" and constitute somewhat of an ultimate arbiter on all matters related to test development and usage. The importance of the *Standards* for addressing the problems associated with testing Hispanic students cannot be overstated.

What is singularly unique is that the current *Standards* have outdistanced current test technology and testing practices with "Individuals of Differing Linguistic Backgrounds." The gulf between what the *Standards* promulgate and what test developers and test users actually do is very large. Given the directives proposed by the Office for Civil Rights in their "Nondiscrimination in High-Stakes Testing: A Resource Guide" (U.S. Department of Education, draft, December 1999), this gulf may well constitute a denial of substantive due process with Hispanic students and citizens. The following is a historical look at how the current *Standards* evolved with respect to testing "bilingual" individuals. A review of the OCR Guidelines then follows.

THE STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING

The first set of these *Standards* appeared in 1966, even though both the American Educational Research Association and the American Psychological Association had addressed the issues attendant to achievement and psychological/diagnostic testing in two prior, separate documents respectively in the mid-1950s.

The 1966 *Standards for Educational and Psychological Tests and Manuals* (American Psychological Association, American Educational Research Association, National Council on Measurement in Education, 1966) had one overriding goal: "the essential principle underlying this document is that a test manual should carry information sufficient to enable any qualified user to make sound judgments regarding the usefulness and interpretation of the test" (pg. 2). Behind this "essential principle" was the recognition "that tests are used in arriving at decisions which may have great influence on the ultimate welfare of the persons tested, on educational points of view and practices, and on development and utilization of human resources" (pg. 1). Interestingly, in the entire 36 pages of text, there are only occasional references to general demographic variables that should be addressed by test manuals. There are only two instances when these references vaguely touch on linguistic and cultural diversity. In both instances they are not prescribed as ESSENTIAL:

C5.5. If the validity of the test is likely to be different for subsamples that can be identified when the test is given, the manual should report the results for each subsample separately or should report that no differences were found. VERY DESIREABLE (pg. 20)

D.2.21. [concerning the psychometric index of reliability] **Demographic information, such as distributions of the subjects with respect to age, sex, socioeconomic level, intellectual level, employment status or history, and minority group membership should be given in the test manual.** DESIREABLE (pg. 28)

In 1974, the new edition of the *Standards for Educational and Psychological Tests* (American Psychological Association, American Educational Research Association, National Council on Measurement in Education, 1974) paid more attention to the issues of cultural and linguistic diversity. There was recognition that the validity of a test could be attenuated for certain groups and under certain conditions. This acknowledgement was due, in great part, to the impact of federal court cases alleging diagnostic bias in tests. In many school districts, tests produced inflated incidence rates of mental disabilities among Latino and African American student populations. Typically, these inflated incidence rates appeared with greater frequency than in the past.

B1.3. The manual should call attention to marked influences on test scores known to be associated with region, socioeconomic status, race, creed, color, national origin, or sex. Essential

[Comment: Social or cultural factors known to affect performance on the test differentially, administrator errors that are frequently repeated, examiner-examinee differences, and other factors that may result in spurious or unfair test scores should, for example, be clearly and prominently identified in the manual.] (pg. 14)

Of even greater importance, two warnings appeared about the possible impact of tests and testing practices on English-language learners. Standard G2 directed test users to know the research literature on tests and testing particularly with respect to the problems associated with testing individuals with "limited or restricted cultural exposure." Standard G2 suggested that the overrepresentation of African American and Spanish-speaking children "with limited cultural exposure" was caused by test users' lack of knowledge about the limitations of tests when cultural differences existed.

Standards J5, J.5.3, and J.5.3.1. went even further. They recommended that when there were great cultural differences between the test taker and the test's norming sample, the tester should not test ("Essential"). They also set forth an accommodation that has become exceedingly popular: when there are no appropriate tests for a given person or population, the tester should use "a broad-based approach to assessment using as many methods as are available to him. Very Desirable" (pg. 71). What this was interpreted to mean by many was to do more assessments with more tests. The Comment for this Standard elaborated on this.

> [Comment: The standard is to do the best one can. This perhaps includes the use of a test, even though no appropriate normative data are available, simply as a means of finding out how the individual approaches the task of the test. It might include references, extensive interviews, or perhaps some *ad hoc* situational tasks. Efforts to help solve educational or psychological problems should not be abandoned simply because of the absence of an appropriate standardized instrument.] (pg. 71)

When a test or tests are not appropriate, giving more tests simply to see how an individual handles the testing situation is questionable. Data exist showing that in some high-stakes testing situations, giving more tests helps neither the tester (Mehan, Hertweck, & Meihls, 1986) nor the child (Taylor, 1991). When a test is not appropriate, it should not be given if it may hurt an individual or lead to serious negative consequences for the individual. The use of an inappropriate test can only be justified if it has little or no consequences for the individual and if it helps assess system effects on similar individuals. For Hispanic students, the use of inappropriate tests is a national problem with a long history of abuse. As the Comment cited above underscores, there are alternative ways to help solve psychological and educational problems. With Hispanic children, these must be linguistically and culturally appropriate.

In 1985, the Third Edition of the *Standards* was published (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1985). The challenges posed by the multiple aspects of diversity (culture, language, disability, gender, socioeconomic status, and the like) appear throughout the entire document. But the most significant evolution of the *Standards* is Chapter 13, "Testing Linguistic Minorities." In many ways, Chapter 13 was revolutionary.

The text that introduced the seven Standards for this chapter began with the most profound acknowledgment. "For a non-native English speaker or for a speaker of some dialects of English, every test given in English becomes, in part, a language or literacy test" (pg. 73) of English. Basically, this means that for bilinguals who have been exposed to another language, every test, except a test of English language proficiency, contains an unknown, systematic degree of error. Such tests, in effect, are biased because they may not be measuring accurately whatever is being measured. Accordingly, the Standards called for "special attention" to these issues on the part of test development, test use, and test interpretation. It was also recognized that bilingual individuals vary extensively in their functional, academic and literate use of each language separately or simultaneously. Also, cognitive processing in the weaker language is more fragile and can be slower. Language background, in effect, is an important consideration in all aspects of testing and test validity.

With respect to using tests that are in the primary language of bilingual individuals, the *Standards* made several, key pronouncements. Translating a test does not guarantee that the test items will have the same degree of difficulty in the other language. The latter must be empirically established. For example, a straight translation of a second-grade test of reading ability will not necessarily yield a second-grade

reading test in the other language. Tests for determining English language proficiency are vitally important for making educational placement decisions. However, these tests must assess multiple dimensions of linguistic ability. Chapter 13 also made a distinction between "natural" uses of language and more formal, cognitively demanding uses. Because of these "special difficulties" attendant on the use of tests with persons who have not had adequate exposure to the language of the test, it was suggested that **more** testing and observations be done with them. As previously noted, more testing was and is a questionable policy.

Chapter 13 also acknowledged the possible influence of culturally mediated ways of responding to test questions. Elaborated speech may not be congruent with culturally specific ways of speaking to adults. When these factors are ignored the validity of interpretations and recommendations may be questionable and harmful.

Chapter 13 of the 1985 *Standards* was vitally important for the educational and psychological testing of English-language learners in the United States. There were, however, several problems. First, it was not known how well the testing industry and professions would abide by them. Throughout the 1980s and 1990s, the manuals of most tests showed that Chapter 13 of *Standards* was routinely ignored. Also, these Standards ignored several, historical practices and problems. They did not address one of the most widely used techniques for testing English-language learners--the use of interpreters who either translate the test into the primary language on the spot or who help administer a test that has already been translated. They were silent on the apparent inability of tests and test users to differentiate among cultural factors, language proficiency levels, and mental/emotional disabilities. They endorsed a historical solution for what to do when there are no tests ("test more") in spite of the fact that there was no evidence that this worked. In fact, there was evidence to the contrary.

In August of 1999, a new edition of the *Standards* was approved. Chapter 9 is titled "Testing Individuals of Differing Linguistic Backgrounds." Just as in the 1985 *Standards*, the narrative introducing this chapter cautions that, with individuals from diverse linguistic backgrounds, tests that are in English become tests of English ability to a degree that is generally more pronounced than with monolingual English speakers. With individuals of varying levels of bilingualism, tests may fail to measure what they intend to measure. Accordingly, norms developed for monolingual English-speaking populations should either not be used or should be interpreted with the understanding that English language proficiency is a contaminating factor. Precautions regarding processing speed factors are also raised. The chapter suggests accommodations should be undertaken with English-language learners. It also notes that cultural factors can affect test scores, so attention should be paid to these factors. The problem with this part of the narrative in Chapter 9 is that, in spite of acknowledging the complexities associated with testing bilingual students, the precautions are tenuous and weak.

Chapter 9 repeats many historical caveats about translating tests without conducting norming studies. Back translations are specifically mentioned as being inadequate by themselves. A similar point was made in the 1985 *Standards*. Yet, publications describing and endorsing this process continue to appear (Geisinger,

1994a,b). A translated test is an inappropriate test. The practice should be proscribed nationally.

The 1999 Standards break new ground along several dimensions. They note that several issues raised in Chapter 9 apply to persons with disabilities that affect communication such as deafness and visual impairments. This connection with disability appears to be part of a general trend in addressing the issues related to linguistic diversity. It also appears in a new, important text commissioned by the American Psychological Association (Sandoval, Frisby, Geisinger, Scheuneman, & Grenier, 1998). But, historically, bilingualism all too often has been equated with a handicap. Linking the test accommodations appropriate for bilingual learners with those appropriate for students with disabilities is exceedingly problematic for bilingual children. Accommodations may seem similar, but their use and outcomes may be different for bilingual children (Thurlow, Liu, Erickson, Spicuzza, & El Sawaf, 1996). This is a matter for serious consideration. Being bilingual is not a handicap, but an asset.

The new 1999 *Standards* discuss several types of test accommodations that may have to be done with English-language learners: using only sections of the test that match the linguistic proficiency of the test-taker, changing the test and response formats, administering the test in a different context, and allowing more time for taking the test. Most of these modifications are currently under study. It is difficult to see, however, how these will overcome the well-documented, historical impact of bilingualism on tests. It is difficult to see how testing a student in the wrong language, or testing for content that has not been taught, or testing for cultural material that is not in a Hispanic child's repertoire will be made fair by the changes suggested in these accommodations.

The issue of "equivalence" receives a great deal of attention in the new *Standards*. This refers to several aspects of test-development, use and interpretation, for example: the degree of confidence that a test-user can exercise in determining whether a test score means the same for someone who is unlike the norming population, the equivalence across translated and renormed versions of the same test, and equivalence in psychometric characteristics. As versions of the same test appear in both English and Spanish, this will become a major topic for research and examination. However, some have asserted that the conceptual basis for testing bilingual children in the United States using monolingual norms in both Spanish and English may be flawed (Grosjean, 1989; Valdés & Figueroa, 1994). It is argued that a bilingual child cannot validly be compared against norms for children whose linguistic experience and development is with only one language. Bilingual children need norms derived from bilingual norming samples, controlling for differential levels of linguistic proficiencies. This issue urgently needs empirical studies and calls for an immediate analysis.

A new directive in these *Standards* calls for taking into account a determination of both language dominance and language proficiency. Consideration should be given to the possibility that bilinguals may have "domain-specific" competencies in one or both languages. For example, a bilingual person may have competency in speaking Spanish but not in reading Spanish. It is recommended that an individual's degree and type of bilingualism be understood in order to use test results properly. This directive has

particular relevance for statewide testing programs. Clearly, if achievement measures are interpreted without the degree of linguistic information suggested by the new Standards, test results may not be correctly analyzed or understood.

Extensive attention is given to the process of administering tests and to the possible impact of test-giver variables (culture, bilingualism, gender, time limits, and the use of interpreters). A crucial principle is recommended for testing English-language learners: give them enough time to finish the test and to show what they know and what they can do. This principle should also be applied in statewide testing programs across the United States.

One of the most surprising parts of these new Standards is the attention given to the use of interpreters. Not only are there multiple precautions, there is also a veritable map for how to train and use interpreters. The unfortunate part of this section of the *Standards* is that there is no empirical evidence that even remotely validates any of the procedures for using interpreters. In fact, several, new dissertations are reporting findings to the contrary (e.g., Sánchez-Boyce, 1999).

There are actually 11 Standards in this new chapter, "Testing Individuals of Differing Linguistic Backgrounds." Most are similar to the ones promulgated in the 1985 edition. Because of their vital importance for the testing of Hispanic children, they are each reviewed and critiqued here. In Appendix A, they are reproduced in their entirety.

There are several meta-issues in these new Standards that are very important. Test users are charged with the responsibility of determining when a test may be inappropriate with linguistic minorities because they do not know "the language of the test" (Comment for Standard 9.1). Similarly, test developers are held responsible, plausibly under "legal or regulatory requirements," for collecting evidence of test validity when there is research indicating differential meaning for test scores for a linguistic group. In a break with historical practices, the use of "representative " norming samples for these validity studies is proscribed. **Separate** norming studies specific to a linguistic group are called for (Comment for Standard 9.2).

Test users are also required to use professional judgement in order to determine language proficiencies prior to testing. Then they are to test in either the most proficient language or using both languages in order to assure construct validity (Comment for Standard 9.3). This is a highly ambitious directive that rests in some exceedingly tenuous assumptions: as it applies to Hispanic students, it is assumed that there are equivalent language proficiency tests in Spanish and English, that such equivalent tests can measure the complexity of linguistic proficiency in both languages, that such tests would have universal application among Hispanic Americans in the United States, that variation in linguistic proficiencies can be used to interpret an individual's test score (how does one interpret a score in a language that is 70 percent proficient and another score in a language that is 55 percent proficient?), and that bilingualism is the sum of two languages (in which case language proficiency testing makes some sense) rather than a linguistic unit (in which case linguistic proficiency testing may be of limited use).

Standard 9.4 seems to tacitly accept the validity of "linguistic modifications" of tests that are in English. These may involve changing the test to the test taker's primary language (translating?) or altering the test given in English. There are no justifications for this Standard (9.4) and the historical, empirical literature reviewed in this document argues against modifications such as translations. The Standards (Comment on 9.4) place the responsibility for justifying these modifications on test developers. The current research on test modifications for English learners is not sufficient to warrant the existence of this Standard.

Standard 9.5 addresses the issue of "flagging" a test score when "linguistic modifications" were provided during the testing. The Standard basically suggests that such flagging may be unfair, and not useful if the score from the modified administration is "comparable" to the score on the "nonmodified" administration and if there is "no reasonable basis" for thinking that such modification affects "score comparability." This is a very problematic Standard because of its lack of specificity. Are tests to be administered to linguistic minority individuals with and without modifications to see if there is comparability? What is a "reasonable basis" for determining comparability among scores? More than anything, the problem with this Standard comes from the unknowns related to "linguistic modifications," or what in the literature is known as "test accommodations" for English language learners.

The current, available literature on such accommodations (Abebi, 1999a,b,c; Thurlow, Liu, Erickson, Spicuzza, & El Sawaf, 1996) makes several points. The use of accommodations varies greatly across and within states that provide such test adaptations. The question of who gets accommodations also varies greatly within and across these states. The most popular accommodations in statewide testing programs are: allowing for extra time, using a bilingual dictionary, being tested in a separate room, receiving oral translations of directions, offering multiple testing sessions, answering questions, providing written and oral translations, having words defined, and allowing for students to mark the test booklets (Thurlow, Liu, Erickson, Spicuzza, & El Sawaf, 1996). These researchers also note: "Few accommodations are universally allowed, and further research on the appropriateness and technical adequacy of different types of accommodations would be beneficial" (pg. 13).

Actual, empirical studies of accommodations (Abebi, 1999a,b,c) have produced modest results in improved test scores of bilingual children. This applies to achievement tests that are among the easiest to "accommodate," namely, math tests. In fact, one could argue that the study of accommodations has made its greatest contribution to children who do not need accommodations. Researchers have found that tests often include test language that is needlessly obtuse and immaterial to the construct being measured. Cleaning up such language improves the performance of all test takers.

Research on test accommodations is currently insufficient to support Standard 9.5. Testing bilingual, Hispanic children on an English test with accommodations may not be adequate to remove the high level of "distortion" or the construct-irrelevant error (Kopriva, 1999) implicated in the assessment of bilingual learners since the 1920s. Accommodations, in effect, may prove to be a subterfuge procedure for testing in the

wrong language. Data already exist showing that some of the most popular accommodations--translating and interpreting--simply do not work.

Standard 9.6 requires test developers and users to explicitly address the questions related to test use and interpretation with non-native speakers. This Standard seems to exclude simultaneous bilinguals from such consideration. It is, regrettably, an example of the 1999 Standards' lack of precision about the complexity of bilingualism.

Standard 9.7 establishes a rule for using translated tests. The methods of translating have to be described and so does the evidence for establishing validity and reliability across language groups. This basically asserts, once and for all, the need to go beyond translations before interpreting or using test scores. But this Standard is problematic. Properly translating a test and then establishing its reliability and validity within different Hispanic cultural groups is really only a first step. There is also a need to establish the validity and reliability of the test within different levels of linguistic proficiencies within different Hispanic cultural groups.

This is another example of how the Standards are fundamentally naïve about the linguistic nature of Hispanic populations in the United States. A well-translated, well-normed test will be confronted not by Spanish speakers of one level of proficiency, but by the typical bilingual Spanish- speaking population of this country: culturally diverse **and bilingually** diverse.

Standard 9.8 speaks to the need for concurrence between what a test measures and what a credential or an occupation demands in terms of actual performance on the job. Particular attention is given in this Standard to the equivalence that should exist in the **linguistic** demands of the test and those of the job.

Standard 9.9 requires that tests that are available in two languages provide evidence that each linguistic version is comparable to the other in terms of reliability, validity (particularly construct validity) and other data. Once again, however, this Standard does not address the existential reality of bilinguals in the United States. Tests that are available in two languages have to demonstrate equivalence across a wide spectrum of linguistic abilities.

Standard 9.10 is critical to this chapter as well as to all testing of Hispanic individuals. The measurement of linguistic proficiency should be done across "a range of language features" (pg. 154) and in more than one testing format (such as multiple choice). Language proficiency is a crucial covariate or control measure in much of what Chapter 9 of the new *Standards* proposes as solutions to testing bilingual individuals. Here, the requirement is that language proficiency be measured in multiple ways. This is an important directive, a critically necessary albeit insufficient step in testing bilingual populations. What remains unaddressed is how such multiple measures of linguistic proficiency are to be used for the interpretation of test scores in both the primary and the secondary language and across the language proficiency profiles that exist in Hispanic and other bilingual populations.

Standard 9.11 is on interpreters. It touches on what is probably the most common historical accommodation in the testing of Hispanics. Unfortunately, it is the most problematic Standard in this chapter. As mentioned, there are no data to substantiate the assumption that it is possible to use an interpreter without severely and negatively affecting the standardization requisites, psychometric properties and the interpretation of test scores. The Standard seems to sanction the translation of tests by the interpreter and requires that the tester assume responsibility for the competence of the interpreter when there is no empirically validated model for training interpreters. Also, in most real-life situations, it is the school district or the clinic that is responsible for selecting, training, and assigning interpreters to testing situations. This is a Standard that urgently needs more deliberation and research.

One interesting omission in the new Standards is the historical accommodation that when there are no appropriate tests available, more testing is an implied and acceptable methodology. This is a welcome change. However, this change should be broadly publicized in order to stop the practice of "more testing."

THE OFFICE FOR CIVIL RIGHTS' *RESOURCE GUIDE*

The issue of nondiscriminatory testing took on a unique significance in the federal courts in the early 1970s because of the overrepresentation of minority students in classes for pupils with disabilities. The most current and extensive analysis of nondiscriminatory assessment has been done by the U.S. Office for Civil Rights (U.S. Department of Education, draft, December 1999). Their *Resource Guide*, however, does not just address nondiscriminatory assessment from the perspective of special education diagnosis. It extends the application of nondiscriminatory assessment beyond special education to all "high-stakes testing" and all assessment methods (norm-referenced, criterion-referenced, and alternative testing methods). An important point highlighted by this *Resource Guide* is that nondiscriminatory assessment must be seen as part of the high-standards movement in American education. It must include, *a priori*, equity in the provision of opportunities to learn for all students.

In educational contexts, tests function as measures of system accountability and as measures of current status or prediction for the student. With this in mind, the *Resource Guide* underlines a critical distinction made by the courts between educational and employment testing:

> If tests predict that a person is going to be a poor employee, the employer can legitimately deny the person the job, but if tests suggest that a young child is probably going to be a poor student, a school cannot on that basis alone deny that child the opportunity to improve and develop the academic skills necessary to success in our society. (Larry P. v. Riles, 1984; cited in U.S. Department of Education, draft, December 1999, pg. ii)

The OCR *Resource Guide* is fundamentally an exposition of the "testing and assessment principles that lie at the core of Title VI of the Civil Rights Act of 1964 (Title VI) and Title IX of the Education Amendments of 1972 (Title IX)..." (U.S. Department of

Education, draft, December 1999, pg. i). Two legal theories of test discrimination are presented: disparate impact and disparate treatment.

The analysis of disparate impact concentrates on whether test practices and policies, regardless of neutrality of application, produce "adverse consequences" (pg. iv) with specific racial, gender or national origin groups. The negative consequences described include "granting or denial of benefits or opportunities" (pg. iv). "Educational necessity", for example placement or student designations, is the only exception in this regard. This means that tests must be reliable and valid for their intended educational purpose. Further, tests may not be used under this analysis if they are "not the least discriminatory practical alternative that can serve the educational institution's educational purpose" (pg. iv). There are three criteria that define "disparate impact." First, a test yields disparate results based on race, national origin, or gender. Second, the test has no educational utility. Third, there are no other practical, valid, or reliable alternatives to assess the students.

For Hispanic students, ample evidence exists showing that tests do cause disparate outcomes in high-stakes decisions: high special education representation rates for LEP students in certain categories of disabilities (United States Department of Education, 1993); low representation rates in programs for the Gifted and Talented (Callahan, Hunsaker, Adams, Moore, & Blend, 1995); and low representation rates in higher education (President's Advisory Commission on Educational Excellence for Hispanic Americans, 1996).

For all students, the question of a test's **educational** usefulness can often be answered in the negative with respect to curricular, remedial, or pedagogical decisions **about an individual**. For example, the most test-driven educational document, the special education Individualized Education Program (IEP), has not produced substantive educational usefullness to children with disabilities (Skrtic, 1991). A clear distinction needs to be made here that "educational usefulness" is both an individual as well as a system consideration. It may work for the latter but not the former, particularly when a test score for a bilingual Hispanic child contains systematic error (American Educational Research Association, et al., 1985, Chapter 13; 1999, Chapter 9). The educational usefulness of tests to Hispanic students is an issue that needs serious, empirical consideration. The possible attenuation of tests' predictive validities with Hispanic, bilingual populations augurs badly for most forms of instructional validity or educational utility.

Invalid inferences are highly probable when tests are used on Hispanic children with varying degrees of exposure to a language other than English. The tests measure something other than what they intend to measure. Predictive validity studies that control for language background strongly indicate that psychometric bias is a real possibility in the testing of students from diverse linguistic backgrounds (Figueroa, 1990; Figueroa & García, 1994). There is a great need for large, longitudinal studies on the predictive validity of tests used in educational contexts holding linguistic background and proficiencies as controls. If this type of predictive bias is further substantiated, the legal theory of disparate impact with Hispanic students would be significantly strengthened.

Practical alternatives to these tests include grades, portfolios, and student work products keyed to rubrics. However, the requirement that the measures be reliable psychometrically is somewhat problematic. The history of testing Hispanic students clearly shows that reliability indices are insensitive to linguistic or cultural differences. Also, the requirements for educationally useful alternatives to tests should focus on visual and instructional indices of validity for appropriate consequences since criterion indices of validity with Hispanic students are currently suspect (Figueroa, 1990; Figueroa & García, 1994).

On the matter of cut-off scores, the OCR *Resource Guide* relies on the principle that "the method and rationale for setting the cut score, including the technical analyses, should be presented in a manual or in a report" (American Educational Research Association, et al., 1985, Standard 6.9). The most prevalent use of such scores occurs in colleges and universities with respect to admissions. Yet, institutions of higher learning typically do not provide any technical analyses that would justify a particular cut-off score on educational grounds. More often than not, universities set up cut-off scores without any empirical consideration as to what score differentiates between those who can learn in university settings and those who cannot. Cut-off scores ignore the systematic undereducation of Hispanic students. Whereas institutions of higher learning may see them as indices of merit, for most Hispanic students they are also measures of unequal opportunities to learn at the K-12 level. Cut-off scores are largely responsible for Hispanics' underrepresentation in institutions of higher learning. Further, as the National Council of La Raza reports, California's Proposition 209 banning affirmative action programs in colleges and universities may re-establish the prominence of using cut-off scores in the SAT and GRE exams for admission purposes. The impact of this would only exacerbate an already inequitable situation. Between 1987 and 1997, Hispanic students' SAT scores decreased in relation to white students' (National Council of La Raza, 1998).

The analysis of disparate treatment focuses on whether testing policies or practices are done differently for individuals or groups with distinct racial, national origin, or gender characteristics. Examples of differential treatments would include "being tested under different conditions" (pg. iv) or "whether students with the same test scores are....treated differently by an educational institution" (pg. iv). The OCR *Resource Guide* fails to consider the possibility that, for a Hispanic student from a linguistically or culturally different background, tests administered in English are tests given "under different conditions" (pg. iv) than those for a monolingual, monocultural student. Clearly, given the evidence reviewed here and acknowledged by the *Standards for Educational and Psychological Testing* (American Educational Research Association, et al., 1985; 1999) this is an issue that should be addressed by the U.S. Office for Civil Rights.

The *Resource Guide*'s section titled Equal Opportunity for Limited-English Proficient Students (pg. 9) is quite inadequate in this respect. Some of its suggested accommodations lack any empirical justification (such as "bilingual dictionaries") and may actually attenuate psychometric properties. Similarly, the suggested "Remedies" are

problematic for Hispanic children and youth: test more, revise the test, substitute the test.

Presently, a strong argument can be made that tests produce disparate impacts and that they do constitute a disparate treatment with regard to Hispanic students from diverse linguistic backgrounds. The viability of such arguments should be debated. It should be done on two levels: the legal/policy level and the psychometric/professional level (with respect to consequential validity).

CHAPTER 4: BIAS

In the early part of the 20th century, the discussion of bias in tests and testing focused on two areas: the impact of the "language handicap" experienced by bilingual individuals and the possible misinterpretation of test data to assert genetic differences among groups, particularly with regards to intelligence. But it was not until the 1960s that the problem of test bias became prominent. Tests were linked with tracking and segregation policies in school districts in several court cases. Most notably in *Hobsen v. Hansen* (1967), the pivotal use of academic aptitude tests for tracking African American children in vocational, high school programs was outlawed by the court. The court found that the tests were biased because they could not really measure student learning potential and because they produced the sort of segregation proscribed by *Brown v. Board of Education*.

In *Hobsen v. Hansen* (1967) and subsequent litigation that involved testing practices (*Diana v. California Board of Education*, 1970; *Larry P. v. Riles*, 1979), test bias became linked with the civil rights meaning of bias, discrimination and prejudice. One of the consequences of this linkage was a vigorous response from the testing community in the form of extensive research on the empirical documentation of bias. Studies conducted on racial/ethnic groups across the full spectrum of available tests were fairly unanimous: test bias could not be found (Cleary, Humphreys, Kendrick, & Wesman, 1975) in the multiple indices of reliability (items, factors, alternate forms, retest) and all the various forms of validity (content, criterion, concurrent, construct).

However, a careful examination and interpretation of the research data on Hispanics since the 1920s suggests that there is evidence of bias. Table 1 presents the sources of test bias and the degree of empirical evidence available.

TABLE 1: SOURCES OF TEST BIAS WITH HISPANIC TEST-TAKERS

1) SIGNIFICANT EXPOSURE TO A LANGUAGE OTHER-THAN ENGLISH
   -Extensive documentation

2) PROCESSING SPEED IN THE WEAKER LANGUAGE
   -Extensive documentation

3) USING TRANSLATIONS OF TESTS
   -Extensive documentation

4) DIMINISHED OPPORTUNITY TO LEARN
   -Extensive documentation

5) USING INTERPRETERS DURING TESTING
   -Emerging documentation

6) DECISION-MAKING BASED ON TESTS
   -Emerging documentation

The following is an extended explanation of each source of bias presented in Table 1.

1) SIGNIFICANT EXPOSURE TO A LANGUAGE OTHER-THAN ENGLISH

Since the 1920s, Hispanic children and adults have consistently demonstrated a classic test profile on tests of mental ability and academic achievement: low [English] verbal scores and high non-verbal/math test scores. The depressed verbal scores are directly related to the degree of exposure to Spanish in the home and the community. Its not that the tests are inherently flawed, but rather that they are applied to the wrong population. Tests in English, since they are generally normed on monolingual English-speaking populations, inherently tap a developmental sequence of English proficiency and English literacy. When exposure to English varies in degree across chronological ages (as with simultaneous bilinguals), or in the time of onset (as with sequential bilinguals), or both, tests register this as a subtrahend. Item analysis studies (Sandoval, 1979; Figueroa,1983) clearly show this as a generalized impact throughout the test items rather than as a discrete phenomenon affecting some items and not others. This explains why studies of test bias in test item structures have not found such bias (Cotter & Burke, 1981). It also explains why internal indices of test reliability and stability have also not found bias with Hispanic children and adults (Valdés & Figueroa, 1994). Item difficulty levels are not changed. The total scores are lower, but the test items perform the same way regardless of English proficiency.

The most powerful impact from exposure to Spanish is manifested in one of the most critical functions of tests: prediction. Though empirical data have suggested this from the beginnings of psychometrics, more recent studies have clearly documented that the greater the degree of exposure to Spanish the lower the predictive validity of [English] tests (Gándara, Keogh, Yashioka-Maxwell, 1980; Pilkington, Piersel, & Ponterotto, 1988; Emerling, 1990; Kaufman & Wang, 1992; Stone, 1992; Valdez & Valdez, 1983; Valencia, 1982; Valencia & Rankin, 1988; Figueroa, 1990; Figueroa & Garcia, 1995; Pennock-Roman, 1990).

The clearest example of this comes from the validity study of the *System of Multicultural Pluralistic Assessment* (Mercer, 1979). Mercer developed a battery of tests (WISC-R IQ's) that purported to operationalize the federal requirement of nondiscriminatory testing in special education diagnoses. She normed the tests in 1972 on what is in all likelihood the most random and representative sample of Hispanic children (N=700) in California. A critical aspect of this study was that the children were all judged to be English proficient by the school staff and by those who individually administered the tests. However, the children came from three types of linguistic environments: homes where only Spanish was spoken, where Spanish and English were spoken, and where only English was spoken. In 1982, a predictive validity study of the tests was undertaken (Figueroa & Sassenrath, 1989) on approximately half of the original norming sample. It was found that on the WISC-R IQ's the predictive validity coefficients for the Hispanic children varied in direct proportion to their exposure to Spanish. Figure 1 presents these data.

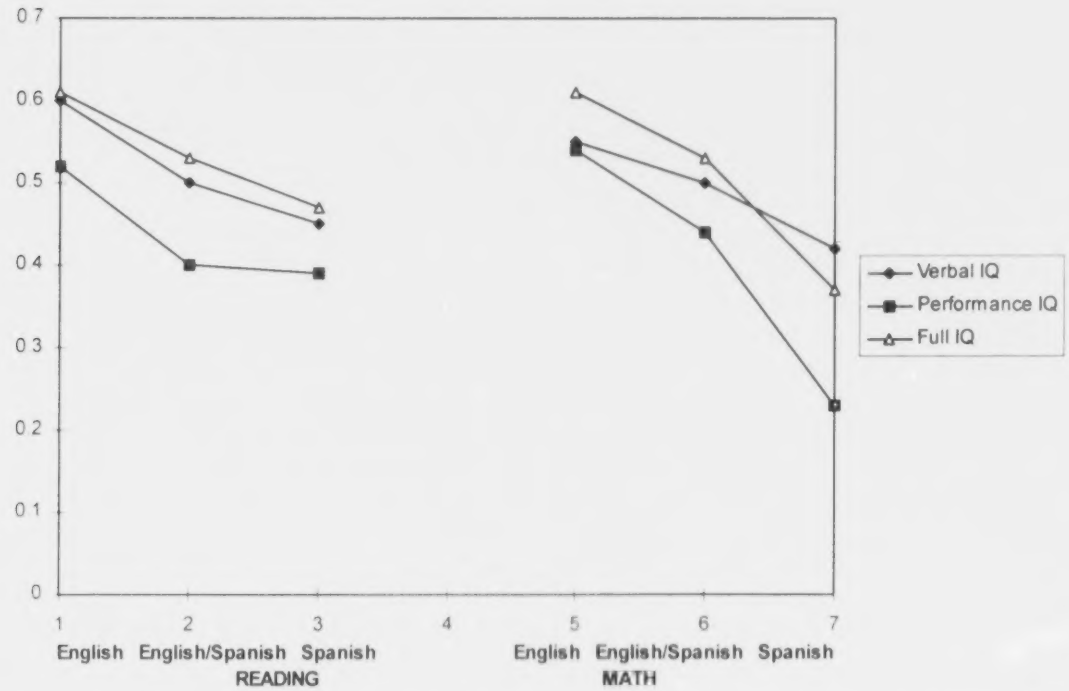


Figure 1: Evidence of psychometric bias in the predictive validity of 1972 IQ's relative to 1982 standardized measures of reading and math achievement for Hispanic children's home language background

As Figure 1 shows, the more Spanish there was in the home in 1972, the less the IQ predicted reading and math achievement in 1982. Ironically, the nonverbal IQ, the historically used solution for measuring mental abilities and aptitudes in bilinguals, proved to be the most hypersensitive to Spanish in the home. Other studies show similar results (Figueroa & Garcia, 1994). However, a definitive predictive validity study using multiple measures of the Hispanic subjects' English proficiency as a control urgently needs to be done. Systematic differences in predictive validity are one of the key signatures of bias in tests.

2) PROCESSING SPEED IN THE WEAKER LANGUAGE

Differences in cognitive processing speed between monolinguals and bilinguals is dramatically demonstrated in the work of Dornic (1979, 1978a, 1978b, 1977) in Europe. Basically, he found that processing information in the weaker language produced consistently slower functioning. Further, the entire process of mentation became progressively more and more vulnerable (to the point of shutting down) when the material was too complex, when the testing situation was too noisy, or when stress levels increased. The importance of these findings for testing that is done under timed, noisy or stressful conditions with Hispanic children merits further research. It should be noted, however, that accommodations providing more time for English learners are already routinely recommended.

These are critical findings whose relevance may extend beyond issues of test fairness or bias. Processing speed and automaticity are crucial requisites for certain academic skills. Currently, for example, the great emphasis being given to fluency in reading as a major indicator of reading proficiency may well be biased or invalid when applied to Hispanic English-language learners.

Research has also continued to document how Hispanic and Japanese bilingual children are better at naming digits backward than monolingual children (Jensen & Inouye, 1980; Figueroa, 1987). Though the typical explanation attributes this to "bilingualism," a more precise explanation may be slower processing in English in children from Spanish-speaking homes. Slower processing may actually be favorable for naming digits backward. Just as likely, slower processing may come from translating material into Spanish and then back into English. In conditions where "slow" is not an impediment, there may be better remembering and possibly better learning (Malakoff & Hakuta, 1991).

3) USING TRANSLATIONS OF TESTS

The 1985 and 1999 Standards place all manner of caveats on translating tests without any re-norming on the target populations for which the translation was developed. This limits the effectiveness of test translations. The rationale for this is quite clear. When a test is normed, the item difficulty levels and the actual norms flow from the responses of the norming sample. Translating a test, no matter how well done, no matter if the translation is then back-translated to English, does not necessarily produce either an equivalent or a useful instrument. But are the caveats in the 1985 and 1999 Standards about translating tests devoid of any empirical proof? Can psychometricians really proceed with elaborate directives on how to translate tests (Geisinger, 1994b) without having to actually re-norm the new, translated test?

In 1982, the Mexican government undertook a renorming of the WISC-R intelligence test in Mexico City (Gómez-Palacio, Padilla, & Roll, 1983). They began with a straight translation of the verbal test items as well as all the instructions. They also added extra items to most of the verbal subtests in order to make these more aligned to Mexican children's opportunities-to-learn in both their public schools and their communities. When the norming was finished, approximately 80 percent of the items that were translated from the English remained. The critical questions in this discussion are: What happened to the test items? Did they remain in the same location as when they were in the translation and in the English version?

Not only did the item sequences, or the degree of difficulty that the children experienced with each word, change; they did so in a most unusual manner. In the first half of the vocabulary subtest, the items were generally easier for Mexican children. In the second half, there were more items that were harder. Overall, Figure 1 clearly shows that equivalence of tests merely through translation does not work. This refutes the practices associated with just translating a test. The new translated test, in all likelihood,

will not be equivalent to the first. The Mexico City data indicate that translating a test, no matter how well done, is a biased procedure. It produces an instrument with unknown psychometric characteristics. It precludes any useful decisions based on the scores.

4) Tests are based on a set of critical assumptions. It is assumed that the individual being tested has had similar experiences as the individuals who generate the test norms. It is concomitantly assumed that there is general equivalence in the opportunities that individuals have had to learn the content in the test, the linguistic genre of the test, and the demands of tests. Intelligence and achievement tests are particularly dependent on meeting these assumptions since without them the attributions to intellectual or academic abilities cannot be reasonably made.

In the case of achievement testing, however, it is possible to determine from multiple research sources when the opportunity to learn for a population is not even or not fair (Orfield & Yun, 1999; Moreno, 1999). In this case, the test scores may reflect or even assess opportunity to learn. They become measures of system accountability. Such scores, however, may not reflect the learning ability of the individual nor his/her potential for learning. In effect, any high-stakes decisions based on scores that do not meet the assumptions of some equivalence in opportunity to learn can be unfair and invalid. Such may be the case in statewide testing programs when the tests are administered in English to English-language learners. The scores themselves become, in unknown degrees, measures of opportunity to learn (and English language proficiency), not of individual achievement and certainly not of future academic achievement.

In 1982, the National Academy of Sciences broke tradition with American psychology (Heller, Holzman, & Messick, 1982). It suggested that individual differences in academic achievement may not be the primary source of score differences. It recommended that **before** a child is tested for special education diagnosis, his or her present instructional setting be evaluated for its validity, effectiveness, and delivery. Similar considerations should be taken into account when testing Hispanic children from diverse cultural, linguistic and schooling communities in the United States.

5) The 1985 *Standards for Educational and Psychological Testing* was criticized (Valdés & Figueroa, 1994) for not addressing one of the most commonly used practices in the testing of bilingual individuals, that is, the use of interpreters. The 1999 Standards finally discuss this practice. But they also endorse it and set about describing the who, how and what for using an interpreter during testing. The unfortunate aspect of this is that there is no empirical evidence in the testing literature that can attest to the procedural equivalence of this process to doing testing in one language or to any scientific documentation that the process actually works reasonably well.

In fact, the few doctoral studies that have recently been done investigating the use of interpreters conclude the opposite (DuFon, 1991; Sánchez-Boyce, 1999). Sánchez-Boyce's dissertation (1999) describes the actual process of using an interpreter during individualized testing sessions for special education placement of Hispanic students as chaotic, erratic, and fairly devoid of any standardization

procedures. She documents how the conclusions reached from such testing sessions are really socially constructed and have no bearing on what the child can or cannot actually do. Research is urgently needed in this widespread practice to either refute or validate these investigations. If the latter turns out to be the case, the 1999 *Standards* should be revised relative to their endorsement of the use of interpreters and it should be done before the next comprehensive draft appears around 2010.

6) The actual process of making decisions on the basis of test scores administered in either English or Spanish has never received much attention. Recently, however, Sandoval (1998) has heuristically taken findings from the research literature on decision-making theory and attempted to apply them to the testing situation where the subject is from a different cultural and linguistic background. Table 3 presents the multiple set of factors that a test-user must engage or consider when testing a Hispanic subject.

TABLE 2: FACTORS TO CONSIDER IN MAKING DIAGNOSTIC DECISIONS WHERE ISSUES OF LINGUISTIC AND CULTURAL DIVERSITY APPLY

1. REPRESENTATIVE BIAS
- Ignoring the prevalence of a behavior in a given population
- Reaching decisions on the basis of a limited sample of observations
- Failing to take into account the fact that some scores will be off by chance
- Making premature casual inferences on the basis of correlations that are not generalizable, coherent, consistent, robust, or reversible

2. CONFIRMATORY BIAS
- Bias from what is expected or what is stereotypic

3. AVAILABILITY BIAS
- Bias that comes from vivid, recent data

4. INTELLECTUAL BIAS
- Bias due to intellectual limitations or cognitive overload due to high degrees of complexity (for example, interactions among cultural, linguistic, and opportunity-to-learn variables)

Assuming that even two of these factors are robust in doing testing and diagnostic work with Hispanic populations, two implications arise: Can anyone with cultural and linguistic backgrounds that are different from the student being tested actually do the task? and, Is it possible to train individuals to effectively use these parameters in making test/diagnostic decisions? As noted earlier, decision-making has never been adequately studied with multilingual and multicultural populations. The distinct possibility exists that this has been and continues to be an inadequately studied source of test bias with Hispanic populations.

The content of this chapter provides empirical evidence that tests used with Hispanic students show evidence of bias. Comprehensive, longitudinal investigations on this question should be commissioned. The impact of Hispanic culture and Spanish language proficiency levels on the predictive, consequential, and/or instructional validity indices of tests should be determined.

CHAPTER 5. EDUCATIONAL ACCOUNTABILITY

For all Hispanic Americans, achievement tests are one of the most important sources of information affecting their lives and communities. The emphasis on higher academic standards for American schools has brought with it an unparalleled degree of concern about system and student accountability. Achievement tests supposedly fulfill this goal better than any other indicator. The precarious aspect of this is that achievement tests are among the most difficult measurement instruments to develop and interpret particularly when they are given in group situations in order to compare academic gains across states, school districts and schools (Cronbach, Linn, Brennan, & Haertel, 1997). These are fragile instruments that suffer from low reliabilities and that all too often assess not just what has been learned but also degrees of English-language proficiency, cultural differences, socioeconomic status in the home and community, quality of past and present pedagogy, and the economic advantage or disadvantage of school districts.

Historically, achievement tests have nearly always described a chronic pattern of underachievement for Hispanic students of all ages. Reynolds (1933) called attention to the one to two-standard deviation differences in achievement test scores between Anglo and Hispanic American populations in the Southwest. In the Coleman Report (Coleman et al, 1966), the academic levels of Mexican American and Puerto Rican children continued to show a one to two standard deviation deficit compared to white children. In the 1970s the U.S. Commission on Civil Rights (1971a, 1971b, 1972, 1973, 1974) documented a similar level of underachievement for Mexican American children. These reports also found that the education of Mexican American children differed significantly from that of white children: fewer questions from their teachers, less reinforcement for their classroom responses, poorer schools, no reflection of their ethnic/cultural background in the curricula, and no Mexican American models in the teaching, administrative, or counseling staffs. Other data in the 1980s and 1990s also continued to show evidence of comprehensive, national levels of underachievement among Hispanic children and youth (National Commission on Secondary Education for Hispanics, 1984a,b; Arias, 1986; Valencia, 1991; President's Advisory Commission on Educational Excellence for Hispanic Americans, 1996; National Council of La Raza, 1998; Laosa, 1998; Moreno, 1999).

Although there were some modest gains between the NAEP achievement scores of Hispanic students across the country between 1988 and 1994, particularly in math and science, the overall picture is one of decline. The achievement gap between white and Hispanic students continues to increase. Key indicators (such as lower enrollment in preschool programs than white students, higher Hispanic enrollment below modal grade, underrepresentation in gifted and talented programs, increased enrollment in segregated schools, a 103 percent increase in suspension rates, a growing digital divide, an increase in the drop-out rates between white and Hispanic students) show that Hispanic students continue to have different educational experiences than their white counterparts in the public schools of the United States.

Yet achievement tests have always operated under the belief that the public schools provide similar educational experiences and opportunities for all students. In 1975, this belief was explicitly noted in a report from the American Psychological Association (Cleary, Humphreys, Kendrick, & Wesman, 1975):

> It is recognized that three assumptions are basic to this report. The first assumption is acceptance of a single society, heterogeneous though it may be, rather than a divided one. The second assumption is that few radical changes are expected in curriculum content or methodology of instruction in our educational establishment.....The third assumption accepts the importance of evaluation in education.

The assumption of homogeneity of "opportunity to learn" in the public schools of the United States is never really mentioned, but it is implied. With all achievement testing, this is a necessary condition that must be met prior to any interpretation about a student's or a group's level of academic achievement. Recently, Orfield & Yun (1999) have documented a national trend towards more segregation of Hispanic students in the public schools of the United States. As has always been the case, segregated schools typically have fewer financial resources, the most inexperienced teachers, and the lowest academic achievement levels. In effect, the Hispanic student does not receive the same curricula or pedagogy or resources as do those students in affluent or middle-income school districts. The issue of differential "opportunities to learn" for Hispanic children in the public schools of the United States during the last 100 years is incontrovertible.

According to the legal analysis on "Nondiscrimination in High-Stakes Testing" authored by the U.S. Office for Civil Rights, this condition of limited opportunity to learn establishes a claim of substantive due process violation related to achievement testing because "the students were not taught the material on which the tests were based" (U.S. Department of Education, draft, December 1999, pg. 2). This is a national claim and one that needs to be addressed, particularly in the current climate of setting progressively higher and higher academic standards without a concomitant resolve to equalize opportunities to learn.

NEW INITIATIVES

The 1990s produced an unprecedented degree of attention to the measurement of academic achievement in Hispanic students. For example, considerable legislative work was focused on including English language learners in all large-scale achievement testing programs (Goals 2000, Educate America Act, P.L. 103-227; the Perkins Act, P.L. 98-524; Improving America's Schools Act, P.L. 103-328). Three initiatives, however, are particularly noteworthy because of their potential impact on national policy and discussion on this matter: the new *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999), the National Research Council's report *Improving Schooling for Language-Minority Children* (August & Hakuta, 1997), and *Grading the Nation's Report Card*, by the National Research Council. Each of these

addresses the challenges involved in testing the academic achievement of English language learners.

In the new *Standards for Educational and Psychological Testing* (American Educational Research Association, et al., 1999), there is a particularly powerful caveat noted with respect to achievement tests: it is not known how much they become measures of English language proficiency when they are used with individuals whose primary language is not just English. ["..among non-native speakers of the language of the test, one may not know whether a test designed to measure primarily academic achievement becomes in whole or in part a measure of proficiency in the language of the test" (pg. 9-4)]. Across the United States, however, the test scores of English-language learners are never described in these terms nor is the possible degree of test bias or error recognized.

Statewide testing programs of academic achievement should include statewide standardized measures of English language proficiency capable of measuring multiple dimensions of this competence. As the 1999 testing Standards suggest, a determination should be made first about which language is dominant. Then, the degree of proficiency in the dominant language should be measured along dimensions such as reading, writing, comprehension, grammar, pronunciation, and communicative competence. There should be a clear understanding, however, that even with this type of comprehensive language proficiency assessment, knowledge in certain domains may be missed by achievement tests. The new Standards therefore recommend doing academic testing in both languages even when proficiency in English is established.

In 1997, the National Research Council, through its Committee on Developing a Research Agenda on the Education of Limited-English-Proficient and Bilingual Students, ostensibly summarized the current knowledge base on testing English language learners in its report *Improving Schooling for Language-Minority Children* (August & Hakuta, 1997). Four contemporary concerns are addressed: the measurement and use of students' L1 and L2 linguistic proficiencies in school districts across the United States, the use of tests with bilingual students for entrance and exit criteria from educational programs (including Title I and special education), the impact of L1 on the validity and reliability of tests, and the measurement of academic achievement particularly in standards-driven educational contexts.

The report sets forth the following research agenda for the assessment of second language learners (August & Hakuta, 1997, pgs. 113-134):

1) Given that current tests tend to measure only discrete linguistic features, the assessment of linguistic proficiencies in L1 and L2 needs to be aligned with current research on how language acquisition occurs in children in bilingual communities.

2) Because different dimensions of English are required by different academic subjects and in different grades, it is necessary to determine how to use

multifaceted measurements of English proficiency to validly predict success in English-only classrooms.

3) Research is needed on how to measure knowledge in academic subjects. Specifically: Does testing in English underestimate academic knowledge if English proficiency is limited? Does lack of familiarity with "test language" affect academic test scores even when the tests are given in the primary language? Is it better to test in English or in the primary language when subject matter has been given only in English? What is the impact on test scores from varying levels of native language proficiency, years of schooling in English, and difficulty of academic content?

4) Studies need to be done on how English learners take tests, specifically, how test demands, test format and test language (such as instructions) affect scores. These studies also need to answer the questions: when can an English learner validly and reliably take a test in English and when does an English learner need test modifications or accommodations? Further, what is the impact of such modifications on test reliability and validity?

5) Studies are needed to determine how rater or scorer error can be reduced in "open-ended or performance-based" (pg. 130) assessments when the test-taker's English proficiency can influence scoring decisions.

6) The standards and accountability reform movements in education call for content, performance and opportunity-to-learn benchmarks (McLaughlin & Shepard, 1995) throughout schools, school districts and federal programs. Research is needed on how to operationalize these for English language learners. Specifically: Can indicators of subject matter competence and English proficiency development be produced for English language learners? How can the progress of English language learners (ELLs) be gauged within school district standards and on indices of academic achievement? If nonstandard assessments are used with ELLs, how can these be included within state and district accountability measures? What is the operational meaning of "yearly progress" for ELLs given the possibility that ELLs "may take more time to meet ...standards" (pg. 127)? Finally, given the fact that there are few data on effective pedagogical, curricular, or contextual conditions for the schooling of ELLs, how can opportunity-to-learn standards be operationalized for them?

There is one area missing in this research agenda. In spite of the fact that there is text (pgs. 124-125) in the report on meeting the assessment needs of English language learners referred for special education testing, the report fails to heed its own voice. Because there are no assessment instruments that can differentiate between linguistic/cultural differences and disabilities, research is needed on how to operationalize the "nondiscriminatory assessment" provisions of federal special education laws, as these apply to Hispanic, English language learners.

By and large, the research agenda proposed by the National Research Council should be endorsed and funded. Its importance is twofold. First, it points to work that is vitally needed now. Second, it highlights the elementary stage that the country is in with respect to measuring the academic achievement levels of English language learners.

The National Research Council has also evaluated the efforts of the National Assessment of Educational Progress (NAEP) in measuring the academic achievement of all students in the United States (Pellegrino, Jones, & Mitchell, 1999. NAEP is one of the more problematic test-developers and test-users with respect to the academic achievement testing of Hispanic students. Up to the 1990s, NAEP, for example, did not even have reliable procedures for identifying the ethnic background of Hispanic students (Rivera, 1986; Rivera & Pennock-Roman, 1987; Baratz-Snowden, Pollack, & Rock, 1988). The manner in which the students were chosen for inclusion in NAEP testing was not random and systematically excluded English language learners. Even when NAEP attempted to address the complexities associated with testing Hispanic children, its efforts were so flawed that it precluded any meaningful interpretation of test scores (e.g., Baratz-Snowden, Rock, Pollack, & Wilder, 1988).

In the most recent report on NAEP, the National Research Council (Pellegrino, Jones, & Mitchell, 1999) pays particular attention to the assessment of English language learners. Asserting that NAEP and other assessment programs have made many efforts to accommodate the special needs of English language learners (Olson & Goldstein, 1997), this report highlights the efforts of the Puerto Rico Assessment of Educational Progress, a unique Spanish language translation and accommodation of NAEP mathematics tests. After the administration of NAEP in Puerto Rico, several key findings about the complexities associated with transporting American tests across languages were made: translations often fail (Anderson & Olson, 1996), and item response theory analyses yield noncomparable scales for English and Spanish versions of the test (Olson & Goldstein, 1997). In another study (Anderson, Jenkins, & Miller, 1996), similar conclusions were reached with respect to the translation of other NAEP tests: "the translated versions of the assessment are not parallel in measurement properties to the English version and scores are not comparable" (pg. 31).

In 1995, a field test of the NAEP mathematics test included more English language learners than ever before. In fact, where previously the NAEP instructed schools across the nation on which ELL students to exclude, in this field test the instructions were on how to include ELL students who the school staff thought could actually take the test. Also, the following accommodations were included: more time, more testing sessions, different testing sessions (group and individual), using an interpreter to elaborate on instructions, and test booklets in Spanish. There is some evidence that accommodations do enhance participation, but the explicit effects of such accommodations on test scores are not clear. Recent research from UCLA suggests that the benefits may be marginal (Abedi, 1999a, b, c). Problems also remain with respect to the test technology used to identify and classify ELL students nationally and with respect to the wide variation across states and school districts in the criteria that they use for these purposes (August & Lara, 1996; Valdés & Figueroa, 1994). As the National Research Council acknowledges:

> To date, the dilemmas described ..... have not been resolved. Children potentially in need of native language support are still being assessed at entry level using one of several instruments that many scholars have questioned, and some years later they are tested again using another of such instruments that is in no way comparable to the first. The field is no closer to developing means for assessing whether a child can or cannot function satisfactorily in an all-English program–or participate in all-English large-scale assessments–than it was in 1964. (Pellegrino, Jones, & Mitchell, 1999, pg. 105)

Table 3 presents the "research agenda" suggested by the National Research Council in order to help NAEP cope with the inclusion of English language learners as part of the Nation's Report Card. As will be noted, some of these (such as using translations) are questionable given the historical and scientific experience of the country with such procedures.

TABLE 3: THE NATIONAL RESEARCH COUNCIL'S CONSIDERATIONS FOR INCLUDING ENGLISH LANGUAGE LEARNERS IN THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

- What types of demands do different assessments make on English language learners and how do different types of accommodations help different types of students?
- How do accommodations affect the construct validity of the tests?
- Is it useful and cost effective to try accommodations such as translations?
- Do scores from accommodated administrations have the same scaling properties and can they be reported in the same fashion as for all other students in NAEP?
- Do English language learners have the same opportunity to learn and curricula as non-ELL students?
- What are possible, alternative assessment methods for ELLs?

(Pellegrino, Jones, & Mitchell, 1999, pg. 110-111)

Table 4 presents the major conclusions and recommendations of the National Research Council for NAEP's testing of English language learners.

TABLE 4: THE NATIONAL RESEARCH COUNCIL'S MAJOR CONCLUSIONS AND RECOMMENDATIONS FOR TESTING ENGLISH LANGUAGE LEARNERS ON THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS (NAEP)

CONCLUSIONS

Conclusion 3A. The participation and accommodation of .... English-language learners are necessary if NAEP results are to be representative of the nation's students. There is currently a paucity of interpretable achievement data and accompanying contextual data on the performance and educational needs of these populations.

Conclusion 3B. Enhanced participation of ..... English language learners in NAEP depends on (1) the consistent application of well-defined criteria to identify these students and (2) accurate collection and reporting of information about them.

RECOMMENDATIONS

Recommendation 3A. NAEP should include sufficient numbers of ..... English language learners in the large-scale assessment so that the results are representative of the nation and reliable subgroup information can be reported.

Recommendation 3B. Criteria for identifying ..... English language learners for inclusion in the large-scale survey need to be more clearly defined and consistently applied.

Recommendation 3C. For students who cannot participate in NAEP's standard large-scale surveys, appropriate, alternative methods should be devised for the ongoing collection of data on their achievement, educational opportunities, and instructional experiences.

Recommendation 3D. In order to accomplish the committee's recommendations, the NAEP program should investigate the following:

XI. Methods for appropriately assessing, providing accommodations, and reporting on the achievements of .... English language learners, and

XII. Effects of changes in inclusion criteria and accommodation trends in achievement results."

(Pellegrino, Jones, & Mitchell, 1999, pg. 112-113)

Table 4 presents an exceedingly modest set of recommendations for English language learners and for Hispanic children. Some of these recommendations go back nearly 70 years (Sánchez, 1934; Reynolds, 1933). It is time for a comprehensive set of action plans to make large-scale, educational accountability systems, such as NAEP, relevant and useful for the educational present and future of Hispanic children. The points made in Table 4 are good **starting** points. But there are other, more complex issues that need to be asked and answered. For example, assuming that it is possible to measure students' multidimensional levels of proficiency in both languages, the challenge remains as to what to do with language proficiency scores. Should they be used to generate expectancy scores? Should they be used to adjust achievement scores to compensate for error? Should they alter cut-offs for eligibility, detention or promotion purposes? Research is needed to establish the function of such scores for interpreting the academic achievement of Hispanic, bilingual individuals with varying levels of acculturation and from the multiple ethnic and cultural backgrounds of Hispanic Americans.

CHAPTER 6. DIAGNOSTIC TESTING FOR SPECIAL EDUCATION

Linguistic minority children, for nearly a century, have tended to overpopulate classes for students with mental disabilities. No other area of education is as linked to issues of genetic differences in intelligence as special education. The fundamental question that has plagued this area of American education is why there are so many minority children with low IQs. Race psychologists in the 1920s and 1930s attributed this to genetic inferiority. Apostles of the same doctrine have asserted the same in the 1960s and continue to do so even now.

> On the other hand you look at a lot of kids in the inner cities who have not seen a book by the time they come to kindergarten, and you give them one and they hold it upside down and the wrong way...The language interactions that they've had at home are nil. They have never even heard these sound systems. Are they lousy readers? A lot of them are. Are they genetically predisposed? Some of them are making that combination a tough one to treat. (Reid Lyon, quoted in Taylor, 1998, pg. 192)

In the 1960s, the federal courts entered this debate. The questions posed then were: Why are there so many minority children in classrooms for the mentally retarded? and, Are the tests used to diagnose mental disabilities biased against them? Many of these court cases focused on the possible linguistic bias of IQ tests and on the denial of equal educational opportunities to Hispanic students placed in special *education (Diana v. California Board of Education*, 1970; *José P. v. Ambach*, 1979; *Arreola v. Board of Education*, 1968; *Guadalupe Organization v. Tempe School District*, 1978; *Ruiz v. State Board of Education*, 1971; *Covarrubias v. San Diego Unified School District*, 1972; *Lora v. Board of Education of the City of New York*, 1984). For the most part, the courts ruled in the direction of using non-verbal tests of intelligence, establishing monitoring systems for determining when Hispanic children were overrepresented in classes for students with mental disabilities, and overseeing training programs for staff in order to do nondiscriminatory assessments. This last provision became part of the federal law for special education in 1975 and was repeated in the mandates for the 1997 Individuals with Disabilities Education Act. However, nondiscriminatory assessment remains either an enigma or serious problem for the American Psychological Association. In its latest, major document on testing individuals from diverse backgrounds (Sandoval et al, 1998), nondiscriminatory assessment does not exist.

The sort of testing that is done in order to determine whether a child has disabilities is unique in education. It is really an analog of what a medical doctor does when an individual has serious symptoms. Most children who are tested for special education placement go to the school psychologist with one predominant symptom, poor reading. After the administration of a few or many, many tests, the psychologist does a diagnosis and at an Individualized Education Program meeting recommends either special or general education placement. At that time he/she also prescribes a treatment to "cure" the symptom. But the similarities between a medical doctor and a school psychologist are illusory. The tests that the school psychologist uses have no real power to diagnose and the educational treatments are usually ineffective (Skrtic, 1991).

Most children who are tested for special education have no clear, biological reasons that triggered the testing. Most children are sent to special education testing between the second and fifth grades. Again, the predominant reason for being sent is poor reading. There are four categories of disabilities (Learning Disabilities, mild Mental Retardation, Behavioral Disorders, Speech and Language Problems) that are unique in this entire enterprise because they are suspect. Many researchers have argued that the diagnostic tests used in special education are incapable of differentiating among these four disabilities (Keogh 1990; Lyon, 1996). Also, for many children, particularly from culturally and linguistically diverse backgrounds, these categories may be socially constructed. That is to say, these students could have a disability or their "symptoms" could be due to socioeconomic, cultural, linguistic or poor opportunity-to-learn factors (Rueda & Forness, 1994; Trueba, 1987; Heller, Holtzman, & Messick, 1982). The tests, however, cannot "diagnose" the difference.

For Hispanic children, the Handicapped Minority Research Institutes in Texas and California (Rueda, Figueroa, Mercado, & Cardoza, 1984; Rueda, Cardoza, Mercer, & Carpenter, 1984; García, 1985; Ortiz, 1986; Ortiz & Maldonado-Colón, 1986; Ortiz & Polyzoi, 1986 ; Ortiz & Yates, 1987; Swedo, 1987; Wilikinson & Ortiz, 1986; Willig & Swedo, 1987) documented the unique problems Hispanic children and their families faced when confronted by the testing/diagnostic process in special education. Language proficiency levels of the students were often not considered. Most testing was done in English. The linguistic challenges experienced by English learners were often diagnosed as a disability. Depending on the tests given, a Hispanic child could easily qualify for the Learning Disability or the Communication Handicapped categories. When retested after being in special education, the Hispanic children's IQs decreased. Limited English Proficient students were more likely to be recategorized with another disability. Tests developed and normed for Spanish-speaking children were just as problematic as tests normed on English speakers. If the parents were born outside the United States, there was a greater likelihood that their child would end up in special education. Finally, it was found that diagnostic tests were capricious in their "diagnoses": when the tests were given to an entire class of Hispanic children in general education, 53 percent were found eligible for the Learning Disability program. When "mentally retarded" Hispanic children were tested, 43 percent of them were "diagnosed" as Learning Disabled. It should be noted that many researchers in the area of special education see no problem in the use of psychometric tests with Hispanic or African American children. For them, the issue of nondiscriminatory assessment simply does not apply, neither do linguistic and cultural factors described in the 1985 or 1999 *Testing Standards* (MacMillan, Gresham, & Bocian, 1998).

Special education testing with Hispanic students has very little empirical, research data to support many of its extant practices. The development of diagnostic tests normed on Spanish-speaking populations abroad provides one of the most widespread uses of diagnostic instruments with Hispanic children. But as some have argued:

> The bilingual is NOT the sum of two complete or incomplete monolinguals: rather he or she has a unique and specific linguistic configuration. The coexistence and constant interaction

> of two languages in the bilingual has produced a different but complete linguistic entity.
> (Grosjean, 1989, pg.6)

> When a bilingual individual confronts a monolingual test, developed by monolingual individuals, and standardized and normed on a monolingual population, both the test taker and the test are asked to do something that they cannot. The bilingual test taker cannot perform like the monolingual. The monolingual test cannot "measure" in the other language. (Hakuta & García, 1989; Hakuta, Ferdman, & Díaz, 1986)

> Ironically, single-language tests deceptively measure the "monolingual" part of the bilingual (one or the other of the bilingual's two languages), irrespective of proficiency in that language, and they do so reliably. But these tests fail insofar as they may exclude mental content that is available to the bilingual in the other language, and mental processes and abilities that are the product of bilingualism.
> (Valdés & Figueroa, 1994, pg. 87)

There is an urgent need to determine the diagnostic validity of Spanish language tests normed on monolingual populations and used for diagnostic purposes with U.S. bilingual populations.

Another current practice in special education testing of Hispanic children is to test repeatedly until the right diagnostic profile appears or to conduct elaborate decision-making procedures in order to get to the "real disability." Data exist, independent of issues of cultural and language differences, indicating that the more testing that is done the less likely the "real disability" will appear (Mehan, Hertweck, & Meihls, 1986; Taylor, 1991).

The use of interpreters is widespread in special education testing (Langdon, 1994). As already mentioned, emerging, empirical data on this practice suggest that it should be proscribed. Another common practice in special education testing is the use of translated tests. This practice should also be stopped.

Though most of this chapter has focused on *diagnostic testing*, other forms of assessment are routinely done in special education (such as curriculum-based assessments, functional assessments, ecological assessments, dynamic assessments). The same general principles that have been discussed with respect to linguistic/cultural differences, bias, validity and reliability apply. For the population of Hispanic children and their families, all of these testing practices also need to be considered in light of several questions: Does placement in special education provide any educational benefits for the student? Does any testing really help to promote educational achievement in special education? Many researchers (Skrtic, 1995; Figueroa & Artiles, 1999; Mehan, Hertweck, & Meihls, 1986) would answer both in the negative.

Finally, there is the question of testing for placement in the gifted and talented programs across the United States. The search for a measure of Hispanic intelligence that would give Hispanic students a fair chance at being equitably represented in such programs has been extensive (Bernal & Reyna, 1975; Chambers, Barron, & Sprecher, 1980; Zappia, 1989; Perrine, 1989; Bermúdez & Rakow, 1990; Márquez, Bermúdez & Rakow, 1992; Johnsen, Ryser, & Dougherty, 1993; Sawyer & Márquez, 1993; García,

1994; Maker, 1996; Maker, Nielson, & Rogers, 1994). By and large, however, none of these have succeeded in establishing a national procedure for identifying gifted Hispanic students. Hispanic pupils, accordingly, are very underrepresented in these programs. They will continue to be absent as long as developers and users of tests and eligibility criteria for gifted and talented programs fail to realize that the opportunity-to-learn experiences of Hispanic children in America's public schools are very different and that tests respond to these differences in the form of lower scores. Cultural factors may also complicate this form of assessment. Data suggest that the family contexts in Hispanic homes of highly academically gifted students vary significantly from Anglo, middle class homes. In Hispanic homes, family values take precedence over individualism and bilingualism is prized over monolingualism (Soto, 1988; Fernández & Nielsen, 1986; Hine, 1993).

For the emerging Hispanic community in the United States, there is an overarching question to be asked about the education of their intelligent children: Is it a good idea to isolate the "smartest" and to give only them an enriched, gifted education? or, Is it a better idea to offer this to all students, including those with disabilities? Some research on programs for the gifted in close-knit communities and in schools suggests that the social impact of these programs can be quite negative for all children, their parents and their communities (Margolin, 1994; Sapon-Shevin, 1994).

CHAPTER 7 – OTHER TYPES OF TESTING

Two other broad types of tests are often given to Hispanic students in educational settings. These are personality and vocational tests. Both are used in counseling settings, though the former is also used for special education diagnoses of emotional disturbance or personality disorders. Both, in varying degrees, are affected by the same issues extant in all other forms of assessments with Hispanic students: cultural differences and linguistic backgrounds.

PERSONALITY TESTING

The empirical and professional literature on personality tests, by and large, has never really attended to the issues of cultural differences or bilingualism in the Hispanic community (Malgady, Constantino, & Rogler, 1987; Olmedo, 1981; Bernal, & Castro, 1994). The official "Guidelines for providers of psychological services to ethnic, linguistic, and culturally diverse populations" (American Psychological Association, 1993) acknowledges that there are problems associated with tests that have not been validated for use with minority populations. But there are no serious limitations placed on their use. Similarly, cultural factors in mental health testing and diagnosis receive very little attention in the "Diagnostic and Statistical Manual of Mental Disorders IV" (Dana, 1995).

As a consequence, some psychological providers recommend that tests be used with interpreters, with norms from other countries (such as Spain), or as a spontaneous translation (such as Nieves-Grafals, 1995; Dana, 1995). Yet, there is evidence that Hispanic test-takers can give different meanings to the connotative, emotive vocabulary that is often used in such tests (Brizuela,1975; Díaz-Guerrero, 1988; González-Reigoza, 1976). Similarly, there is research on how they express affective material in ways that are quite different in Spanish than in English (González, 1978; Ruiz, 1975). Also, data exist on how Hispanic clients are rated differently on personality variables when they are evaluated in English versus when they are evaluated in Spanish (Grand, Marcos, Freedman, & Barroso, 1977; Westermeyer, 1987; Edgerton & Karno, 1971).

A contemporary approach to personality testing involves the parallel use of tests of acculturation. Research on one such test, the "Acculturation Rating Scale for Mexican Americans" (ARSMA I & II) (Cuellar, Harris, & Jasso, 1980; Cuellar, Arnold, & Maldonado, 1995) has proven to be of some importance both in terms of heuristics and improvements in the personality testing of Hispanics (Velásquez & Callahan, 1992). It has been demonstrated, for example, that on the most widely used test of personality, the MMPI, Hispanic Americans whose cultural orientation is "traditional" show more "pathology." These outcomes are basically the result of differential (cultural) treatment. The ARMA studies show this (Dana, 1995) in a particularly "emic" way, though other research similarly confirms that differential cultural impacts persist even in the MMPI-2 (Whitworth & McBlaine, 1993; Whitworth & Unterbrink, 1994).

The use of acculturation scales, such as the ARMA I and II, is usually limited to a specific subgroup of Hispanics. Currently, acculturation scales are more abundant for Cuban American and Mexican American populations (Marin, 1992). Given the limited representation of the scales, their use is limited. Also the use of such scales remains optional. No attempt is being made to psychometrically incorporate such measures into tests such as the MMPI. Admittedly, however, using measures of acculturation or sociocultural variation to "correct" score bias has not fared well in the past. The *Sociocultural Scales of the System of Multicultural Pluralistic Assessment* (SOMPA) (Mercer, 1979), for example, do produce a new IQ estimate based on the degree of distance that a Hispanic child's family exhibits from middle class, white families. But the new IQ proved to be neither a better predictor nor a better "corrector" (Figueroa & Sassenrath, 1989; Valdés & Figueroa, 1994).

There is formal resistance to norming a test such as the MMPI **within** Hispanic populations. This is a reluctance that cuts across all types of tests currently in use. As Dana (1995) has suggested, the press for group-specific norms may have to come from studies showing the error or mistake rates that mono-normative tests produce in personality diagnoses with Hispanics. In many ways these data are already available in all other areas of testing. Most tests do produce negative, differential impacts with Hispanic students.

The present state of testing Hispanics on personality tests relies on a series of questionable practices: making testers "culturally competent," doing "corrections" on the tests, promoting ideographic, tester interpretations based on measures of acculturation. There is very little, actual research on how to do diagnostic personality work with Hispanic children and youth (Cervantes & Arroyo, 1994). As a consequence, recommendations to clinicians, even when they are very well crafted (such as Cervantes & Arroyo, 1994), remain anecdotal and of unknown generalizability, utility and validity. As is the case with most tests used in the United States, new tests specifically made appropriate for the Hispanic populations' bilingual, multicultural status are needed. Preliminary efforts in this regard (Costantino, Malgady, Rogler, & Tsui, 1988) have been fruitful and deserve more research and development. A sea change in attitude about testing distinctly bilingual/multicultural populations is needed within the testing community in the United States. Dana (1995) notes:

> The [current] repertoire of standard tests emerged in an era when a "melting pot" conception of acculturation was in vogue and new immigrants were expected to assume a relatively homogeneous identity after three generations in this country. This expectation did not occur uniformly even for descendents of European immigrants. Now that diversity instead of homogenization has become the hallmark of American society, professional acts and technologies must reflect this societal change. (Dana, 1995, pg. 314)

OCCUPATIONAL INTEREST TESTS

Helping a student choose a career and plan a requisite program of academic preparation are critical counseling functions. Often, this process begins with the administration of interest inventories as early as elementary school. A fundamental

assumption behind these tests is that students in society have somewhat equivalent levels of cultural capital. That is, they have a relatively realistic notion about how societal systems function and what it takes to negotiate entrance into such systems. In terms of careers and jobs, cultural capital means knowing what defines a particular area of occupational interest and what levels are possible within a chosen occupation. Cultural capital in this context also means knowing how to negotiate entrance into systems that help produce the requisite competencies and the desired job or career. For Hispanic students, recent research (Stanton-Salazár, 1997; Stanton-Salazár & Dornbusch, 1995) suggests that, in fact, cultural capital is not easily accessed by them.

Research on the use and impact of occupational interest tests with Hispanic students is neither elaborate nor elegant. There are not many studies and those that do exist do not control for either bilingualism or acculturation in terms of cultural capital.

On the other hand, the related area of employment testing is uniquely optimistic about its fairness, validity and virtual universality (Ramos, 1992) for use with Hispanic adults. Some of this comes from predictive validity studies that show that there are no differences in the way tests function with Hispanic job applicants. The problem here, however, is that these tests are very poor predictors for everybody and that the effects of bilingualism on predictive validity remains unknown. Also, a central element discovered in the employment testing of Hispanics is that so much of the success in the tests and in the jobs depends on educational background.

CHAPTER 8 – SUMMARY AND RECOMMENDATIONS

The testing of Hispanic children has not made much progress in the 20th century. The areas where there has been quite a bit of progress is in the empirical documentation of the impact of bilingualism on test scores and on the development of policies and caveats associated with the testing of Hispanic individuals. However, there has not been much progress in test development and technology in any area of testing with respect to these students.

On the basis of what has been reviewed, seven options appear to exist concerning the measurement of Hispanic students' abilities, achievements, personality, and occupational interests: 1) tests can be administered in English using what are basically monocultural norms, 2) testers can be given "cultural training" so that they can interpret the tests in ways that appear to be more valid, 3) accommodations in the tests and the testing situation can be provided, 4) a moratorium on the use of individual test scores for any high-stakes assessment can be put in place until research sorts out the complex issues associated with testing Hispanic students, 5) tests can be used for holding systems legally and politically accountable for the educational decisions that adversely impact Hispanic students as manifested in differential, negative outcomes, 6) Hispanic-specific local norms can be developed in order to compare students with similar cultural, linguistic, and scholastic experiences, and 7) school systems and opportunities to learn are made equitable for Hispanic children across the United States, thereby meeting the crucial assumption of tests about experiential homogeneity. At present, only the first three are viable and in use. None of these three, however, can demonstrate that they are free of significant degrees of bias, unfairness, or denial of substantive due process.

The fourth option has been suggested (Valdés & Figueroa, 1994) but has received virtually no support. The fifth option has not really been tried in the last decade, but it remains a plausible response to political attacks, such as California's propositions 227 and 209, that are already inflicting harm and damage on Hispanic children and that can be documented by the tests' ability to measure contextual effects. In Kern County in California, for example, the school board has decreed that Hispanic children must learn English in three months and then receive their education in English. The impact of this decision will become manifest in the tests administered in English.

The sixth option may well be the most immediately relevant for both test developers and the Hispanic communities in the United States. But there is a great deal of opposition from both political and professional interests. Ethnic/linguistic norms will provide comparisons among children with generally homogeneous experiences and background in local communities. But, they arouse suspicions about a "divided" society. They may also be seen as sources of reverse discrimination. In employment testing, the courts and Congress have refused to accept group-specific norming precisely because of issues related to reverse discrimination (Sireci & Geisinger, 1998). Ironically, the intellectual community has not been so reluctant. The National Academy of Sciences recommended this as a solution to the bias that results from employment tests among job applicants with differential opportunities-to-learn (Hartigan & Wigdor, 1989).

Education, however, has always occupied a different status with the courts. This may also apply with regard to testing in educational contexts. The issue of group norms in all aspects of schooling should be studied and debated. Certainly, the 1999 *Standards for Educational and Psychological Testing* may already have sided with this option. There are clear mandates there that validity needs to be grounded within linguistic groups when research indicates that test scores are affected by language background (Comment for Standard 9.2).

The seventh is the best option; but if history is any indicator, it is the one most likely to take multi-generations to accomplish. It is also the option that best explains why tests are such a failure for Hispanic communities. The primary problem with tests is not the tests. It is the educational context in which they are developed, used, and studied. The historical and contemporary data have clearly documented that, in the United States, public education has not worked for Hispanic children. Tests help and perpetuate much of the dysfunction that Hispanic children get in schools.

The one positive conclusion that can be drawn from the review presented in this document is that the testing community is finally beginning to realize that the problems with testing Hispanic students are far more complex than ever imagined and that they are potentially irremediable in the *status quo* (Heller, Holtzman, & Messick, 1982; American Educational Research Association, et al., 1985, 1999; Pellegrino, Jones, & Mitchell, 1999; August & Hakuta, 1997; Sandoval et al., 1998; Heubert & Hauser, 1999). The solution to the problems engendered and embodied in tests resides in changing the educational experiences of Hispanic children.

A compelling example of what this may entail was described by García and Otheguy (1995). They set out to answer four research questions in an ethnographic study of "seven private, but low-tuition, non-elite schools in Dade County, Florida." They were "run by and for Cubans." The parents of the children were predominantly from working class and middle class income levels. They were, in effect, similar to families of Hispanic children in urban school districts. The four research questions were typically those that preoccupy educational researchers about bilingual children in U.S. public schools: Should Spanish be used? How is language dominance measured and used? When do you use English? In which language is reading taught? The authors were unable to answer these research questions. The following are the reasons for this failure.

> When majority educators look at the education of Hispanic children in the United States, they focus on their linguistic deficits....Discussions about the education of these children begin and end with the issue of the English language, or how they lack it, and how best to give it to them.....However, when Hispanic parents and educators in control of the education of their own children think about the educational process, they ask different questions. They ask questions about the way to educate their children, about pedagogy, instructional strategies and teaching methods, about curriculum and materials. We asked them about language, they told us about education..... Spanish naturally belongs in ethnic schools that are controlled, staffed and run by the Hispanic community, so there is no need to question its role in public education.....
>
> Those of us in public education need to learn from these educators that substantive high expectations do matter; that bilingualism and biliteracy are obtainable if

> one holds both children and teachers unequivocally responsible for obtaining them; that initial literacy in two languages is possible and doesn't have to be limited to Spanish; that advanced literacy in two languages is possible and doesn't have to be limited to English; that in U.S. society all children acquire English naturally and that therefore English acquisition should not be the main focus of education; that parents and community do matter for education; that when they are in control....the results are ultimately superior; that the context of a child's home culture is essential..; and that continuity with the intellectual and social climate of the home is of paramount importance if the school is to help children develop and foster their intellectual and social growth.
>
> (García and Otheguy, 1995, pg. 99-100)

The public education of Hispanic children needs to focus on education. It needs to be reformed pre-eminently in terms of local control. Until such time as when the U.S. educational system is locally and proportionally controlled by Hispanic communities and until it achieves a modicum of equity in how it distributes resources, cultural capital, and the application of "high standards" across all school districts, tests and test scores will continue to show massive technical problems of bias, differential treatments and differential outcomes. They will continue to impede the future of Hispanic communities. Tests will "work" when the public education of Hispanic children becomes democratic and effective.

Recommendations

1. The U.S. Department of Education's Office for Civil Rights should undertake a legal analysis of test usage with Hispanic students and individuals, focusing on the dysjuncture that exists between what the *Standards for Educational and Psychological Testing* prescribe and what test users (individual testers, school district testing programs, state testing programs) actually do. Particular attention should be focused on the testing of bilingual individuals.

2. The U.S. Department of Education's Office for Civil Rights needs to determine whether the "disparate treatment" legal analysis under Title VI and Title IX statutes applies to the historical experience of many, if not most, Hispanic students with tests and testing. A compelling, empirical argument can be made that they are tested under different conditions: they are tested with monolingual norms when most of them have varying levels of bilingual status, all of which have left an indelible, if not unerasable, mark on all tests that use English as the main vehicle for eliciting responses; and, their scores show evidence of attenuated predictive validity related directly to their varying levels of exposure to Spanish.

3. Excessive testing should be discouraged. There is a widespread belief that with students for whom current testing technology may not be appropriate, the thing to do is to test them more using many different tests. There is no evidence to support this approach. There are, however, data suggesting that excessive testing does not improve diagnostic decisions (Mehan, Hertweck, & Meihls, 1986), but, rather, that it may negatively affect children (Taylor, 1991).

4. The U.S. Department of Education's Office for Civil Rights needs to determine whether the "disparate impact" legal analysis under Title VI and Title IX statutes applies to the comprehensive and chronic pattern of Hispanic students' underrepresentation in Gifted and Talented programs, their overrepresentation in programs for students with disabilities, and their miniscule presence in institutions of higher learning in the United States. There is, in effect, a clear *pattern* of a disparate impact from testing practices across a wide array of tests used in multiple educational contexts. There is also compelling evidence that there is bias in prediction and that this differentially constricts tests' educational purposes when used with most Hispanic students.

5. Translated tests should not be used. There is very little likelihood that the new translated test will have the same technical properties as the original, and there is a substantial likelihood that the translated test will not work. The practice of translating tests and of using their scores for making decisions about individuals should stop.

6. A clear distinction, if not separation, needs to be drawn between the issues that are significant in meeting the challenges of a disability with those involved in the education of children with two linguistic systems. Recent publications on "diversity" and "test accommodations" are linking the issues relevant to English language learners with those that are meaningful for students with disabilities. One of the great

historical mistakes in American education has been the tendency to perceive bilingualism as a handicap. For example, special education is dedicated to diminishing the impact of a disability. The education of English learners should not be guided by the diminution of an asset such as bilingualism.

7. Tests that purport to have equivalent test versions in English and Spanish need to show empirical evidence that, in fact, there is equivalence. Similarly, research is urgently needed on whether bilingual, Hispanic children in the United States can be validly and fairly compared on Spanish/English tests that relied on monolingual samples to generate monolingual norms in English and Spanish.

8. The use of interpreters should be discouraged, if not proscribed. Interpreters are basically poor substitutes for what should be provided to Hispanic students: culturally knowledgeable, linguistically competent testers from their own communities. As currently envisioned in the 1999 *Standards for Educational and Psychological Testing*, interpreters can be trained and used in testing situations. New data on this practice, however, suggest that the use of interpreters may somewhat destroy comprehensive standardization. Further, in special education, the use of interpreters may lead to invalid inferences and conclusions. The failure to recruit, train and graduate Hispanics in the testing professions cannot be ameliorated by the use of interpreters. This is a practice that may really be a malpractice.

9. It is recommended that the Standards in Chapter 9 for "Testing Individuals of Diverse Linguistic Backgrounds" be analyzed by experts in second-language acquisition, language proficiency testing, and bilingual assessment in order to examine the ambiguities and the assumptions of that chapter. The 1999 *Standards for Educational and Psychological Testing* are problematic in the areas of language proficiency testing and the use of testing accommodations with bilingual subjects.

10. The impact of Hispanic culture and Spanish language proficiency levels on the predictive, consequential, and/or instructional validity indices of tests should be determined. There is empirical evidence that tests used with Hispanic students show evidence of bias. Comprehensive, longitudinal investigations on this question should be commissioned.

11. The U. S. Department of Education's Office for Civil Rights should conduct an analysis of testing practices with Hispanic students throughout the states and by the National Assessment for Educational Progress to determine whether some or all of these do not meet the legal criteria of discrimination under Title VI and Title IX.

12. The research agenda on assessment proposed by the National Research Council's Committee on Developing a Research Agenda on the Education of Limited-English-Proficient and Bilingual Students in its report *Improving Schooling for Language-Minority Children* (August & Hakuta, 1997) should be endorsed and funded.

13. The recommendations of the National Research Council on testing English language learners on NAEP (Table 4) should be adopted, funded and applied. They should

also be broadened to include Hispanic children from all the major ethnic cultural backgrounds. The issues related to cultural factors in achievement testing (such as acculturation, the measurement of acculturation, the use of acculturation levels) should be investigated. It is time for a comprehensive set of action plans to make large-scale, educational accountability systems, such as NAEP, relevant and useful for the educational present and future of Hispanic children.

14. There is an urgent need to determine the diagnostic validity of Spanish language tests normed on monolingual populations and used for diagnostic purposes with U.S. bilingual populations. Diagnostic tests should not be administered to Hispanic students or they should be relegated to a lower status in the decision-making process for special education, or gifted and talented education. Alternatives to the typical battery of diagnostic tests exist, all the way from placing a child in an enriched treatment situation to "diagnosing" their work products.

15. As is the case with most tests used in the United States, new personality tests specifically made appropriate for the Hispanic population's bilingual, multicultural status are needed. There is very little actual research on how to do diagnostic personality work with Hispanic children and youth.

16. It is recommended that research in occupational interest tests be significantly and quickly increased. Given the widespread use of occupational interest tests with elementary and high school students, as well as their possible role in tracking students in academic programs, the lack of research on the use and impact of these measurement instruments on Hispanic children and youth is a major knowledge gap.

17. Extended analyses and debate need to be conducted on whether Hispanic students' test scores should be interpreted primarily within a school district's "normative framework." That is to say, should national or statewide comparisons that are used to determine an individual's eligibility for promotion, graduation, or admission to higher education continue to be made given the current knowledge base on testing Hispanic students? This does not preclude the use of tests to measure the performance of school systems (schools, districts) to determine how well or how poorly they are working. Clearly, however, in those school districts where there is no equality in educational programs and opportunities for Hispanic students, the question of what constitutes a fair, normative comparison needs to be answered.

18. It should be made clear that the starting point for the reform of unfair testing of Hispanic students is not the tests; It is the instructional context. Until there is some semblance in equity of standards, curricula, pedagogy and resources throughout schools, school districts and states, tests will continue to reify the inequality of educational opportunities in the country. Tests will continue to blame the Hispanic student for low scores and will continue to deny him or her promotion, eligibility and opportunity.

REFERENCES

Abedi, J. (1999a). *NCME examining the effectiveness of accommodation on math performance of English Language Learners*. Paper presented at the 1999 AERA conference, Montreal, Canada.

Abedi, J. (1999b). *The impact of students' background characteristics on accomodation results for students with limited English proficiency*. Paper presented at the 1999 AERA Conference, Montreal, Canada.

Abedi, J. (1999c). *NAEP math test accommodations for students with limited English proficiency*. Paper presented at the 1999 AERA Conference, Montreal, Canada.

Altus, W.D. (1945). Racial and bi-lingual group differences in predictability and in mean aptitude test scores in an Army special training center. *Psychological Bulletin, 42*, 310-320.

Altus, W.D. (1949). The Mexican American: The survival of a culture. *The Journal of Social Psychology, 29*, 211-220.

American Educational Research Association, American Psychological Association, The National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, The National Council on Measurement in Education. (1974). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, The National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Psychological Association (1993). Guidelines for providers of psychological services to ethnic, linguistic, and culturally diverse populations. *American Psychologist, 48*, 45-48.

Ammons, R.B. & Aguero, A. (1950). The Full-Range Picture Vocabulary Test: VII. Results for a Spanish American school-age population. *Journal of Social Psychology, 32*, 3-10.

Anderson, N.E., Jenkins, F.F. & Miller, K.E. (1996). *NAEP inclusion criteria and testing accommodations: Findings from the NAEP 1995 field test in mathematics*. Princeton, NJ: Educational Testing Service.

Anderson, N.E. & Olson, J. (1996). *Puerto Rico Assessment of Educational Progress: 1994 PRAEP Technical Report*. Princeton, NJ: Educational Testing Service.

Arias, B. (Ed.). (1986). The education of Hispanic Americans: A challenge for the future. *American Journal of Education, 95*.

Arreola v. Santa Ana Board of Education (Orange County, California). No. 160-577. (1968).

Arsenian, S. (1945). Bilingualism in the post-war world. *Psychological Bulletin, 42*, 65-86.

August, D. & Hakuta, K. (Eds.). (1997). *Improving schooling for language-minority children: A research agenda*. Washington, DC: National Academy Press.

August, D. & Lara J. (1996). *Systemic reform and limited English proficient students*. Washington, D.C.: Council of Chief State School Officers.

Baratz-Snowden, J., Pollack, J. & Rock, D. (1988). *Quality of responses of selected items on NAEP special study student survey*. Princeton, NJ: National Assessment of Educational Progress.

Baratz-Snowden, J., Rock, D., Pollack, J., & Wilder, G. (1988). *The educational progress of language minority children: Findings from the 1985-86 special study*. Princeton, NJ: Educational Testing Service.

Bell, R. (1935). *Public school education of second generation Japanese in California*. Stanford, CA: Stanford University Press.

Bermudez, A.B. & Rakow, S.J. (1990). Analyzing teachers' perceptions of identification procedures for gifted and talented Hispanic limited English-proficient students at-risk. *Educational Issues of Language Minority Students, 7*, 21-34.

Bernal, E.M. Jr. & Reyna, J. (1975). Analysis and identification of giftedness in Mexican American children: A pilot study. In B.O. Boston (Ed), *A resource manual of information on educating the gifted and talented*. Reston, VA: Council for Exceptional Children.

Bernal, M.E. & Castro, F.G. (1994). Are clinical psychologists prepared for service and research with ethnic minorities? Report of a decade of progress. *American Psychologist, 49*, 797-805.

Brigham, C.C. (1922). *A study of American intelligence*. Princeton, NJ: Princeton University Press.

Brigham, C.C. (1930). Intelligence tests of immigrant groups. *Psychological Review, 37*, 158-165.

Brizuela, C.S. (1975). Semantic differential responses of bilinguals in Argentina. *Dissertation Abstracts International, 36*(12-B), 6439. University Microfilms No. 76-13, 514).

Brown, G.L. (1922). Intelligence as related to nationality. *Journal of Educational Research, 5*, 324-327.

Brown v. Board of Education. 347 U.S. 483 (1954).

Callahan, C.M., Hunsaker, S.L., Adams, C.M., Moore, S.D. & Blend, L.C. (1995). *Instruments used in the identification of gifted and talented students*. Charlottesville, VA: The National Research Center on the Gifted and Talented, Research Monograph 95130.

Cebollero, P.A. (1936). Reactions of Puerto Rican children in New York City to psychological tests. An analysis of the study by Armstrong, Achilles, and Sacks of the same name. San Juan, P.R.: The Puerto Rico School Review, pt. 11.

Cervantes, R.C. & Arroyo, W. (1994). DSM-IV: Implications for Hispanic children and adolescents. *Hispanic Journal of Behavioral Sciences, 16*, 8-27.

Chambers, J.A., Barron, F. & Sprecher, J.W. (1980). Identifying gifted Mexican American students. *Gifted Child Quarterly, 24*, 123-128.

Cleary, T.A., Humphrey, L.G., Kendrick, S.A. & Wesman, A. (1975). Educational uses of tests with disadvantaged students. *American Psychologist, 15*, 15-40.

Coleman, J.S., et al. (1966). Equality of educational opportunity. Washington, D.C.: U.S. Office of Education.

Colvin, S.S. (1921). Intelligence and its measurement: A symposium (IV). *Journal of Educational Psychology, 12*, 136-139.

Costantino, G., Malgady, R.G., Rogler, L.H. & Tsui, E.C. (1988). Discriminant analysis of clinical outpatients and public school children by TEMAS: A Thematic Apperception Test for Hispanics and Blacks. *Journal of Personality Assessment, 52*, 670-678.

Cotter, D.E. & Berk, R.A. (1981). *Item bias in the WISC-R using Black, White, and Hispanic learning disabled children*. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA.

Covarrubias v. San Diego Unified School District. No. 70-394-T, San Diego, CA. (1972).

Cronbach, L.J., Linn, R.L., Brennan, R.L. & Haertel, E.H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational Psychological Measurement, 57*, 373-399.

Cuellar, I., Arnold, B. & Maldonado, R. (1995). Acculturation Rating Scale for Mexican Americans-II: A revision of the original ARSMA scale. *Hispanic Journal of Behavioral Sciences, 17*, 275-304.

Cuellar, I., Harris, L.C. & Jasso, R. (1980). An acculturation scale for Mexican American normal and clinical populations. *Hispanic Journal of Behavioral Sciences, 2*, 197-217.

Dana, R.H. (1995). Culturally competent MMPI assessment of Hispanic populations. *Hispanic Journal of Behavioral Sciences, 17*, 305-319.

Darcy, N.T. (1946). The effect of bilingualism upon the measurement of the intelligence of children of pre-school age. *Journal of Educational Psychology, 38*, 21-44.

Darsie, M.L. (1926). The mental capacity of American-born Japanese children. *Comparative Psychology Monographs, 3*, 1-89.

Davenport, E.L. (1932). The intelligence quotients of Mexican and non-Mexican siblings. *School and Society, 36*, 304-306.

Diana v California State Board of Education. No. C-70-37, United States District Court of Northern California. (1970).

Díaz-Guerrero, R. (1988). *Psicologia del Mexicano* [Psychology of the Mexican] (4th ed.). Mexico: Editorial Trillas.

Dornic, S. (1977). *Information processing and bilingualism*. Department of Psychology, University of Stockholm.

Dornic, S. (1978a). *Noise and language dominance*. Department of Psychology, University of Stockholm.

Dornic, S. (1978b). The bilingual's performance: Language dominance, stress, and individual differences. In D. Gerver & H. Sinaiko (Eds.), *Language and Interpretation and Communication*. New York: Plenum Press.

Dornic, S. (1979). Information processing in bilinguals: Some selected issues. *Psychological Research, 40*, 329-348.

DuFon, M.A. (1991). Politeness in interpreted and non-interpreted IEP (Individualized Education Program) conferences with Hispanic Americans. (Doctoral dissertation, University of Hawaii, 1991). *UMI Dissertation Services*, 1345950.

Edgerton, R.B. & Karno, M. (1971). Mexican-American bilingualism and the perception of mental illness. *Archives of General Psychiatry, 24*, 286-290.

Emerling, F. (1990). An investigation of test bias in two nonverbal cognitive measures for two ethnic groups. *Journal of Psychoeducational Assessment, 3*, 233-244.

Feingold, G.A. (1924). Intelligence of the first generation of immigrant groups. *Journal of Educational Psychology, 15*, 65-82.

Fernández, R.M. & Nielsen, F. (1986). Bilingualism and Hispanic scholastic achievement: Some baseline results. *Social Science Research, 15*, 43-70.

Figueroa, R.A. (1983). Test bias and Hispanic children. *Journal of Special Education, 17*, 431-440.

Figueroa, R.A. (1987). *Special education assessment of Hispanic pupils in California: Looking ahead to the 1990s*. Sacramento: California State Department of Education, Office of Special Education.

Figueroa, R.A. (1990). Assessment of linguistic minority group children. In C.R. Reynolds and R.W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Vol 1. Intelligence and achievement*. New York: Guilford.

Figueroa, R.A. & Artiles, A. (1999). Disproportionate minority placement in special education programs: Old problem, new explanations. In A. Tashakkori & S.H. Ochoa (Eds.), *Readings on Equal Education: Volume 16, Education of Hispanics in the United States: Politics, Policies, and Outcomes* (pgs. 93-118). New York: AMS Press.

Figueroa, R.A. & Garcia, E. (1994). Issues in testing students from culturally and linguistically diverse backgrounds. *Multicultural Education*. Fall, 1994, pgs. 10-19.

Figueroa, R.A. & Sassenrath, J.M. (1989). A longitudinal study of the predictive validity of the System of Multicultural Pluralistic Assessment (SOMPA). *Psychology in the Schools, 26*, 5-19.

Gándara, P., Keogh, B.K., & Yoshioka-Maxwell, B. (1980). Predicting academic performance of Anglo and Mexican American Kindergarten children. *Psychology in the Schools*, 17, 174-177.

Garcia, J.H. (1994). Nonstandardized instruments for the assessment of Mexican-American children for gifted/talented programs. In S.B. Garcia (Ed), *Addressing Cultural and Linguistic Diversity in Special Education: Issues and trends* (pp 46-57). Reston, VA: The Council for Exceptional Children.

Garcia, O. & Otheguy, R. (1995). The bilingual education of Cuban American children in Dade County's ethnic schools. In O.Garcia & C. Baker (Eds). *Policy and practice in bilingual education: A reader extending the foundations* (pgs. 93-102). Clevedon, England: Multilingual Matters.

Garcia, S.B. (1985). Characteristics of limited English proficient Hispanic students served in programs for the learning disabled: Implications for policy, practice and research (Part 1). *Bilingual Special Education Newsletter*, pp1-5.

Garth, T.R., (1920). Racial differences in mental fatigue. *Journal of Applied Psychology, 4*, 235-244.

Garth, T.R. (1928). A study of the intelligence and achievement of full-blooded Indians. *Journal of Applied Psychology, 12*, 511-516.

Geisinger, K.F. (1992). Fairness and selected psychometric issues in the psychological testing of Hispanics. In K.F. Geisinger (Ed.), *Psychological testing of Hispanics* (pgs. 17-42). Washington, DC: American Psychological Association.

Geisinger, K.F. (Ed.). (1992). *Psychological testing of Hispanics*. Washington, DC: American Psychological Association.

Geisinger, K.F. (1994a). Psychometric issues in testing students with disabilities. *Applied Measurement in education, 7*, 121-140.

Geisinger, K.F. (1994b). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*, 304-312.

Gómez-Palacio, M.M., Padilla, E.R. & Roll, S. (1983). *Escala de Inteligencia para Nivel Escolar Weschler* [Wechsler Intelligence Scale for Children-Revised]. Mexico City: El Manual Moderna, S.A. de C.V.

González, J.R. (1978). Language factors affecting treatment of bilingual schizophrenics. *Psychiatric Annuals, 8*, 68-70.

González-Reigoza, F. (1976). The anxiety-arousing effect of taboo words in bilinguals. In C.D. Spielberger & R. Diaz-Guerrero (Eds.), *Cross-cultural anxiety* (pp. 89-105). Washington, DC: Hemisphere.

Gould, S. J. (1981). *The mismeasure of man*. New York: Norton.

Grand, S., Marcos, L., Freedman, N. & Barroso, F. (1977). Relation of psychopathology and bilingualism to kinesthetic aspects of interview behavior in schizophrenia. *Journal of Abnormal Psychology, 86*(5), 492-500.

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language, 36*, 3-15.

Guadalupe Organization v. Tempe Elementary School District. 587 F.2d 1022, U.S. Dist. Court of Arizona. (1978).

Hakuta, K., Ferdman, B.M. & Diaz, R.M. (1986). *Bilingualism and cognitive development: Three perspectives and methodological implications*. Los Angeles: University of California, Los Angeles. Center for Language Education and Research.

Hakuta, K. & Garcia, E.E. (1989). Bilingualism and education. *American Psychologist, 44*, 374-379.

Hartigan, J.A. & Wigdor, A.K. (Eds.1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. Washington, D.C.: National Academy Press.

Heller, K.A., Holtzman, W.H., & Messick, S. (1982). *Placing children in special education: A strategy for equity*. Washington, DC: National Academy Press.

Heubert, J.P. & Hauser, R.M. (1999). *High-stakes testing for tracking, promotion, and graduation*. Washington, D.C.: National Academy Press.

Hine, C.Y. (1993). *The home environment of gifted Puerto Rican children: Family factors which support high achievement*. Presented at the annual National Research Symposium on Limited English Proficient Student Issues, Washington, D.C.

Hobson v Hansen. (1967). 269 F. Supp. 401 (D.D.C.).

Holtzman, W.H., Díaz-Guerrero, R., & Swartz, J.D. (1975). *Personality development in two cultures: A cross-cultural longitudinal study of school children in Mexico and the United States*. Austin: University of Texas Press.

Hung-Hsia, H. (1929). The mentality of the Chinese and Japanese. *Journal of Applied Psychology, 13*, 9-31.

Jensen, A.R. & Inouye, A. R (1980). Level I and Level II abilities in Asian, White, and Black children. *Intelligence, 4*, 41-49.

Johnsen, S.K., Ryser, G., & Dougherty, E. (1993). The validity of product portfolios in the identification of gifted students. *Gifted International: A Talent Development Journal, 8*(1), 40-43.

Johnson, L.W. (1938). A comparison of vocabularies of Anglo-American and Spanish-American high school pupils. *Journal of Educational Psychology, 29*, 135-144.

José P. v. Ambach, No.C-270 (E.D.N.Y. 1979).

Kaufman, A.S. & Wang, J-J. (1992). Gender, race and education differences on the K-BIT at ages 4 to 90 years. *Journal of Psychoeducational Assessment, 10*, 219-229.

Keogh, B. K. (1990). Narrowing the gap between policy and practice. *Exceptional Children*, 57, 186-190.

Koch, H.L., & Simmons, R. (1926). A study of the test performance of American, Mexican and Negro children. *Psychology Monographs, 35*, 1-116.

Kopriva, R.J. (1999). A conceptual framework for the valid and comparable measurement of all students. Washington, D. C.: Council of Chief State School Officers.

Langdon, H.W. (1994). *The interpreter-translator process in the educational setting: A resource manual*. Sacramento, CA: Resources in Special Education, California Department of Education.

Laosa, L. M. (1998). School segregation of children who migrate to the United States from Puerto Rico. Princeton, NJ: Educational Testing Service.

Larry P. v. Riles. 343 F. Supp. 1306 (N.D. Cal. 1972) *affr* 502 F.2d 963 (9th Cir. 1974); 495 F. Supp. 926 (N.D. Cal. 1979); appeal docketed, No. 80-4027 (9th Cir., Jan. 17, 1980).

Lester, O.D. (1929). Performance tests and foreign children. *Journal of Educational Psychology, 20*, 303-309.

Lora v. Board of Education of the City of New York, 587 F. Supp. 1592 (E.D.N.Y. 1984).

Luh, C.W. & Wy, T.M. (1931). A comparative study of the intelligence of Chinese children on the Pintner performance and the Binet tests. *Journal of Social Psychology, 2*, 402-408.

Lyon, G.R. (1996). Learning disabilities. *The Future of Children: Special Education for Students with Disabilities, 6*(1), 54-76. Los Altos, CA: Center for the Future of Children, David and Lucille Packard Foundation.

MacMillan, D.L., Gresham, F.M. & Bocian, K.M. (1998). Discrepancy between definitions of learning disabilities and school practices: An empirical investigation. Journal of *Learning Disabilities, 31*, 314-326.

Maker, C.J. (1996). Identification of gifted minority students: A national problem, needed changes, and a promising solution. *The Gifted Child Quarterly, 40*, 41-50.

Maker, C.J., Nielson R. & Rogers L. (1994). Giftedness, diversity and problem-solving. *Teaching Exceptional Children, 27*, 4-19.

Malakoff, M. & Hakuta, K. (1991). Translation skill and metalinguistic awareness in bilinguals. In E. Bialystock (Ed.), *Language processing in bilingual children*. Cambridge, MA: Cambridge University Press.

Malgady, R., Constantino, G. & Rogler, L. (1984). Development of a thematic apperception test (TEMAS) for urban Hispanic children. *Journal of Consulting and Clinical Psychology, 52*(6), 986-996.

Manuel, H.T. (1935). *Spanish and English editions of the Stanford-Binet in relation to the abilities of Mexican children* (University of Texas Bulletin No. 3532). Austin: University of Texas.

Margolin, L. (1994). *Goodness personified: The emergence of gifted children*. New York: Aldine de Gruyter.

Marin, G. (1992). Issues in the measurement of acculturation among Hispanics. In K.F. Geisinger (Ed.), *Psychological testing of Hispanics* (pgs. 235-252). Washington, DC: American Psychological Association.

Márquez, J.A., Bermúdez, A.B. & Rakow, S.J. (1992). Incorporating community perceptions in the identification of gifted and talented Hispanic students. *The Journal of Educational Issues of Language Minority Students, 10*, 117-127.

McLaughlin, M. W. & Shepard, L.A. (1995). *Improving education through standards-based reform. A report of the National of Education Panel on standards-based education reform*. Stanford, CA: National Academy of Education.

Mehan, H., Hertweck, H. & Meihls, J.L. (1986). *Handicapping the handicapped*. Palo Alto, CA: Stanford University Press.

Mercer, J.R. (1979). *The system of multicultural pluralistic assessment: Technical manual*. New York: Psychological Corporation.

Mitchell, A.J. (1937). The effect of bilingualism on the measurement of intelligence. *Elementary School Journal, 38*, 29-37.

Moreno, J.F. (Ed.)(1999). The elusive quest for equality. Cambridge, MA: Harvard Educational Review.

National Commission on Secondary Education for Hispanics. (1984a). *"Make Something Happen": Hispanic and urban high school reform. V.I.* Washington, DC: Hispanic Policy Development Project.

National Commission on Secondary Education for Hispanics. (1984b). *"Make Something Happen:" Hispanics and urban high school reform, V.II.* Washington, DC: The Hispanic Policy Development Project Inc.

National Council of La Raza. (1998). *Latino education, status and prospects*. Washington, DC: National Council of La Raza.

Neiser, U., Bodoo, G., Bouchard, T.J., Boykin, A.W., Brody, N., Ceci, S.J., Halpern, D.F., Loehlin, J.C., Perloff, R., Sternberg, R.J. & Urbina, S. (1996). Intelligence: Knowns and unknowns [Task Force Report of the American Psychological Association]. *American Psychologist, 51*, 77-101.

Nieves-Grafals, S. (1995). Psychological testing as a diagnostic and therapeutic tool in the treatment of traumatized Latin American and African Refugees. *Cultural Diversity and Mental Health, 1,* 19-27.

Olmedo, E.L. (1981). Testing linguistic minorities. *American Psychologist, 36,* 1078-1085.

Olson, J.F. & Goldstein, A.A. (1997). The inclusion of students with disabilities and limited English proficiency in large-scale assessments: A summary of a recent progress. Washington, DC: U.S. Department of Education.

Orfield, G. & Yun, J.T. (1999). *Resegregation in American schools.* The Civil Rights Project: Harvard University.

Ortiz, A.A. (1986). Characteristics of limited English proficient Hispanic students served in programs for the learning disabled: Implications for policy and practice (Part II). *Bilingual Special Education Newsletter,* pp. 1-5.

Ortiz, A.A. & Maldonado-Colón, E. (1986). Recognizing learning disabilities in bilingual children: How to lessen inappropriate referral of language minority students to special education. *Journal of Reading, Writing, and Learning Disabilities International, 43*(1), 47-56.

Ortiz, A.A. & Polyzoi, E. (1986). *Characteristics of limited English-proficient Hispanic students served in programs for the learning disabled: Implications for policy and practice.* Austin: University of Texas. (ERIC Document Reproduction Service No. ED 267 597)

Ortiz, A.A. & Yates, J.R. (1987). *Characteristics of learning disabled, mentally retarded, and speech-language handicapped Hispanic students at initial evaluation and re-evaluation.* Unpublished manuscript, University of Texas at Austin.

Paschal, F.C. & Sullivan, L.R. (1925). Racial influences in the mental and physical development of Mexican children. *Comparative Psychology Monographs, 3,* 1-76.

Pellegrino, J.W., Jones, L.R., & Mitchell, K.J. (Eds.) (1999). Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress. Washington, DC: National Academy Press.

Pennock-Roman, M. (1990). *Test validity and language background: A study of Hispanic American students at six universities.* New York: College Entrance Examination Board.

Perrine, J. (1989). The efficacy of situational identification of gifted Hispanics in East Los Angeles through nurturance that capitalizes on their culture. In C.J. Maker & S.W. Schiever (Eds.), *Critical issues in gifted education: Vol. 2. Defensible programs for cultural and ethnic minorities.* Austin, TX: Pro-Ed Publishers. (pp.3-18).

Pilkington, C., Piersel, W., & Ponterotto, J. (1988). Home language as a predictor of first grade achievement for Anglo- and Mexican-American children. *Contemporary Educational Psychology, 13*(1), 1-14.

Pintner, R. & Arsenian, S. (1937). The relation of bilingualism to verbal intelligence and school adjustment. *Journal of Educational Research, 31,* 255-263.

Pintner, R. & Keller, R. (1922). Intelligence tests of foreign children. *Journal of Educational Psychology, 13,* 214-222.

Pratt, H.G. (1929). Some conclusions from a comparison of school achievement of certain racial groups. *Journal of Educational Psychology, 20,* 661-668.

President's Advisory Commission on Educational Excellence for Hispanic Americans. (1996). *Our nation on the fault line: Hispanic American education*. Washington, DC: President's Advisory Commission on Educational Excellence for Hispanic Americans.

Ramos, R.A. (1992). Testing and assessment of Hispanics for occupational and management positions: A developmental needs analysis. In K.F. Geisinger (Ed.), *Psychological testing of Hispanics* (pgs. 173-194). Washington, DC: American Psychological Association.

Reynolds, A. (1933). *The education of Spanish-speaking children in five southwestern states* (1933, Bulletin No.11). Washington, DC: U.S. Department of the Interior.

Rivera, C. (1986). *The national assessment of educational progress: Issues and concerns for the assessment of Hispanic students*. Chicago: Spencer Foundation.

Rivera, C. & Pennock-Roman, M. (1987). *Issues in race/ethnicity identification procedures in the national assessment of educational progress, Part I: A comparison of observer reports and self-identification*. Princeton, NJ: Educational Testing Service.

Rueda, R., Cardoza, D., Mercer, J.R., & Carpenter, L. (1984). *An examination of special education decision-making with Hispanic first-time referrals in large urban school districts*. Las Alamitos, CA: Southwest Regional Library.

Rueda, R., Figueroa, R., Mercado, P. & Cardoza, D. (1984). *Performance of Hispanic educable mentally retarded learning disabled and nonclassified students on the WISC-RM, SOMPA and S-KABC*. Los Alamitos, CA: Southwest Regional Laboratory.

Rueda, R. S. & Forness, S.R. (1994). Childhood depression: Ethnic and cultural issues in special education. In R.L. Peterson & S.I. Jordan (Eds.), *Multicultural issues in the education of students with behavioral disorders* (pgs. 40-62). Cambridge, MA: Brookline.

Ruiz, E.J. (1975). Influence of bilingualism on communication groups. International *Journal of Group Psychotherapy, 25*, 391-395.

Ruiz v. State Board of Education. C.A. No. 218394 (Super. Ct. Cal, Sacramento County, 1971).

Saer, D.J. (1923). The effect of bilingualism on intelligence. *British Journal of Psychology, 14*, 25-38.

Sánchez, G.I. (1934). Bilingualism and Mental Measures: A word of caution. *Journal of Applied Psychology, 18*, 765-772.

Sánchez-Boyce, M. (1999).The social construction of meaning in interpreted events: Use of foreign language interpreters in education. Doctoral dissertation, University of California at Davis.

Sandoval, J. (1979). The WISC-R and internal evidence of test bias with minority groups. *Journal of Consulting and Clinical Psychology, 47*, 919-927.

Sandoval, J. (1998). Critical thinking in test interpretation. In , J. Sandoval, C.L. Frisby, K.F. Geisinger, J.D. Scheuneman & J.R. Grenier, (Eds.). *Test interpretation and diversity: Achieving equity in assessment* (pgs. 31-50). Washington, D.C.: American Psychological Association.

Sandoval, J., Frisby, C.L., Geisinger, K.F., Scheuneman, J.D., & Grenier, J.R. (1998). *Test interpretation and diversity: Achieving equity in assessment*. Washington, DC: American Psychological Association.

Sapon-Shevin, M. (1994). *Playing favorites: Gifted education and the disruption of community*. New York: State University Press.

Sawyer, C.B. & Márquez, J.A. (1993). Discrimination against LEP students in gifted and talented classes. *The Journal of Educational Issues of Language Minority Students, 12*, 143-149.

Sireci, S.G. & Geisinger, K.F. (1998). Equity issues in employment testing. In J. Sandoval, C.L. Frisby, K.F. Geisinger, J.D. Scheuneman, & J.R. Grenier (Eds.). *Test interpretation and diversity: Achieving equity in assessment* (pgs. 105-140). Washington, DC: American Psychological Association.

Skrtic, R. (1991). The special education paradox: Equity as the way to excellence. *Harvard Educational Review, 61*, 148-186.

Skrtic, T. M. (Ed.) (1995). Disability and democracy. New York: Teachers College Press.

Smith, M.E. (1942). The effect of bilingual background on college and aptitude scores and grade point ratios earned by students at the University of Hawaii. *Journal of Educational Psychology, 33*, 356-364.

Smith, M.E. (1949). Measurement of vocabularies of young bilingual children in both of the languages used. *Journal of Genetic Psychology, 74*, 305-310.

Smith, M.E. (1957). Word variety as a measure of bilingualism in preschool children. *Journal of Genetic Psychology, 90*, 143-150.

Soto, L.D. (1988). The home environment of higher and lower achieving Puerto Rican children. *Hispanic Journal of Behavioral Sciences, 10*, 161-167.

Stanton-Salazar, R.D. (1997). A social capital framework for understanding the socialization of racial minority children and youth. *Harvard Educational Review, 67*, 1-34.

Stanton-Salazar, R.D. & Dornbusch, S.M. (1995). Social capital and the reproduction of inequality; Information networks among Mexican-origin high school students. *Sociology of Education, 68*, 116-135.

Stone, B.J. (1992). Prediction of achievement by Asian-American and White children. *Journal of School Psychology, 30*, 91-99.

Swedo, J. (1987). Effective teaching strategies for handicapped limited English-proficient students. *Bilingual Special Education Newsletter*, pp.1-5.

Taylor, D. (1991). *Learning Denied*. Portsmouth, NH: Heinemann.

Taylor, D. (1998). *Beginning to read and the spin doctors of science: The political campaign to change America's mind about how children learn to read*. Urbana, IL: National Council of Teachers of English.

Taylor, R.L. & Richards, S.B. (1991). Patterns of intellectual differences of Black, Hispanic, and White children. *Psychology in the Schools, 28*, 5-8.

Thurlow, M., Liu, K., Erickson, R., Spicuzza, R. & El Sawaf, H. (1996). *Accommodations for students with limited English proficiency: Analysis of Guidelines from states with graduation exams*. Minneapolis, MN: National Center on Educational Outcomes.

Trueba, H.T. (1987). *Success or failure? Learning and the language minority student*. New York: Newbury.

U.S. Commission on Civil Rights. (1971a). *Ethnic isolation of Mexican Americans in the public schools of the Southwest* (Report I: Mexican American Educational Study). Washington, DC: U.S. Commission on Civil Rights.

U.S. Commission on Civil Rights. (1971b). *The unfinished education: Outcomes for minorities in the five southwestern states* (Report II: Mexican American Educational Study). Washington, DC: U.S. Commission on Civil Rights.

U.S. Commission on Civil Rights. (1972). *The excluded student: Educational practices affecting Mexican Americans in the Southwest* (Report III: Mexican American Educational Study). Washington, DC: U.S. Commission on Civil Rights.

U.S. Commission on Civil Rights. (1973). *Teachers and students: Differences in teacher interaction with Mexican-American and Anglo students* (Report V: Mexican American Educational Study). Washington, DC: U.S. Commission on Civil Rights.

U.S. Commission on Civil Rights (1974). *Toward quality education for Mexican-Americans* (Report VI: Mexican-American Educational Study). Washington, DC: U.S. Commission on Civil Rights.

U.S. Department of Education (1993). *Fifteenth annual report to Congress on the implementation of The Education of the Handicapped Act*. Washington, DC: U.S. Department of Education.

U.S. Department of Education. (draft, December 1999). Nondiscrimination in high-stakes testing: A resource guide. Washington, DC: U.S. Department of Education.

Valdés, G. & Figueroa, R.A. (1994). *Bilingualism and testing: A special case of bias*. Norwood, NJ: Ablex Publishing.

Valdez, R.S., & Valdez, C. (1983). Detecting predictive bias: The WISC-R vs. achievement scores of Mexican-American and non-minority students. *Learning Disability Quarterly, 6*(4), 440-447.

Valencia, R.R. (1982). Predicting academic achievement of Mexican-American children: Preliminary analysis of the McCarthy Scales. *Educational and Psychological Measurement, 42*, 1269-1278.

Valencia, R.R. (Ed.). (1991). *Chicano school failure and success: Research and policy agendas for the 1990s* (The Stanford Series on Education and Public Policy). Basingstoke, England: Falmer Press.

Valencia, R.R. & Rankin, R.J. (1983). Concurrent validity and reliability of the Kaufman version of the McCarthy Scales Short Form for a sample of Mexican-American children. *Educational and Psychological Measurement, 43*, 915-925.

Velásquez, R.J. & Callahan, W.J. (1992). Psychological testing of Hispanic Americans in clinical settings: Overview and issues. In K.F. Geisinger (Ed.), *Psychological testing of Hispanics* (pp. 253-265). Washington, DC: American Psychological Association.

Westermeyer, J. (1987). Cultural factors in clinical assessment. *Journal of Consulting and Clinical Psychology, 55*, 471-478.

Wheeler, L. R. (1932). The mental growth of dull Italian children. *The Journal of Applied Psychology, 16*, 650-667.

Whitworth, R.H. & McBlaine, D.D. (1993). Comparison of the MMPI and MMPI-2 administered to Anglo- and Hispanic-American university students. *Journal of Personality Assessment, 61*, 19-27.

Whitworth, R.H. & Unterbrink, C. (1994). Comparison of MMPI-2 clinical and content scales administered to Hispanic and Anglo-Americans. *Hispanic Journal of Behavioral Sciences, 16*, 255-264.

Wilkinson, C.Y. & Ortiz, A.A. (1986). *Characteristics of limited English-proficient and English-proficient learning disabled Hispanic students at initial assessment and at reevaluation.* Austin: University of Texas, Handicapped Minority Research Institute on Language Proficiency.

Willig, A.C. & Swedo, J.J. (1987). *Improving teaching strategies for exceptional Hispanic limited-English proficient students: An exploratory study of task engagement and teaching strategies.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Wood, M.M. (1929). Mental test findings with Armenian, Turkish, Greek and Bulgarian subjects. *Journal of Applied Psychology, 13*, 266-273.

Woodrow, H. (1921). Intelligence and its measurement: A symposium (XI). *Journal of Educational Psychology, 12*, 207-210.

Yoder, D. (1928). Present status of the question of racial differences. *Journal of Educational Psychology, 19*, 463-470.

Zappia, I.A. (1989). Identification of gifted Hispanic students: A multidimensional view. In C.J. Maker & S.W. Schiever (Eds.), *Critical issues in gifted education. Vol 2: Defensible programs for cultural and ethnic minorities* (pp. 19-26). Austin, TX: Pro-Ed.

APPENDIX A

The 1999 Standards for "Testing Individuals of Diverse Linguistic Backgrounds" (Chapter 9)

Standard 9.1 Testing practice should be designed to reduce threats to the reliability and validity of test score inferences that may arise from language differences.

Comment: Some tests are inappropriate for use with individuals whose knowledge of the language of the test is questionable. Assessment methods together with careful professional judgment are required to determine when language differences are relevant. Test users can judge how best to address this standard in a particular testing situation.

Standard 9.2 When credible research evidence reports that test scores differ in meaning across subgroups of linguistically diverse test takers, then to the extent feasible, test developers should collect for each linguistic subgroup studied the same form of validity evidence collected for the examinee population as a whole.

Comment: Linguistic subgroups may be found to differ with respect to appropriateness of test content, the internal structure of their test responses, the relation of their test scores to other variables, or the response processes employed by individual examinees. Any such findings need to receive due consideration in the interpretation and use of scores as well as in test revisions. There may also be legal or regulatory requirements to collect subgroup validity evidence. Not all forms of evidence can be examined separately for members of all linguistic groups. The validity argument may rely on existing research literature, for example, and such literature may not be available for some populations. For some kinds of evidence, separate linguistic subgroup analyses may not be feasible due to the limited number of cases available. Data may sometimes be accumulated so that these analyses can be performed after the test has been in use for a period of time. It is important to note that this standard calls for more than representativeness in the selection of samples used for validation or norming studies. Rather, it calls for separate, parallel analyses of data for members of different linguistic groups, sample sizes permitting. If a test is being used while such data are being collected, then cautionary statements are in order regarding the limitations of interpretations based on test scores.

Standard 9.3 When testing an examinee proficient in two or more languages for which the test is available, the examinee's relative language proficiencies should be determined. The test generally should be administered in the test taker's most proficient language, unless proficiency in the less proficient language is part of the assessment.

Comment: Unless the purpose of the testing is to determine proficiency in a particular language or the level of language proficiency required for the test is a work requirement, test users need to take into account the linguistic

characteristics of examinees who are bilingual or use multiple languages. This may require the sole use of one language or use of multiple languages in order to minimize the introduction of construct-irrelevant components to the measurement process. For example, in educational settings, testing in both the language used in school and the native language of the examinee may be necessary in order to determine the optimal kind of instruction required by the examinee. Professional judgement needs to be used to determine the most appropriate procedures for establishing relative language proficiencies. Such procedures may range from self-identification by examinees through formal proficiency testing.

Standard 9.4 Linguistic modifications recommended by test publishers, as well as the rationale for the modifications, should be described in detail in the test manual.

Comment: Linguistic modifications may be recommended for the original test in the primary language or for an adapted version in a secondary language, or both. In any case, the test manual should provide appropriate information regarding the recommended modifications, their rationales, and the appropriate use of scores obtained using these linguistic modifications.

Standard 9.5 When there is credible evidence of score comparability across regular and modified tests or administrations, no flag should be attached to a score. When such evidence is lacking, specific information about the nature of the modifications should be provided, if permitted by law, to assist test users properly to interpret and act on test scores.

Comment: The inclusion of a flag on a test score where a linguistic modification was provided may conflict with legal and social policy goals promoting fairness in the treatment of individuals of diverse linguistic backgrounds. If a score from a modified administration is comparable to a score from a nonmodified administration, there is no need for a flag. Similarly, if a modification is provided for which there is no reasonable basis for believing that the modification would affect score comparability, there is no need for a flag. Further, reporting practices that use asterisks or other non-specific symbols to indicate that a test's administration has been modified provide little useful information to test users.

Standard 9.6 When a test is recommended for use with linguistically diverse test takers, test developers and publishers should provide the information necessary for appropriate test use and interpretation.

Comment: Test developers should include in test manuals and in instructions for score interpretation explicit statements about the applicability of the test with individuals who are not native speakers of the original language of the test. However, it should be recognized that test developers and publishers seldom will find it feasible to conduct studies specific to the large number of linguistic groups found in certain countries.

Standard 9.7 When a test is translated from one language to another, the methods used in establishing the adequacy of the translation should be described, and empirical and logical evidence should be provided for score reliability and the

validity of the translated test's score inferences for the uses intended in the linguistic groups to be tested.

Comment: For example, if a test is translated into Spanish for use with Mexican, Puerto Rican, Cuban, Central American and Spanish populations, score reliability and the validity of test score inferences should be established with members of each of these groups separately where feasible. In addition, the test translation methods used need to be described in detail.

Standard 9.8 In employment and credentialing testing, the proficiency level required in the language of the test should not exceed that appropriate to the relevant occupation or profession.

Comment: Many occupations and professions require a suitable facility in the language of the test. In such cases, a test that is used as a part of selection, advancement, or credentialing may appropriately reflect that aspect of performance. However, the level of language proficiency required on the test should be no greater than the level needed to meet work requirements. Similarly, the modality in which language proficiency is assessed should be comparable to that on the job. For example, if the job requires only that employees understand verbal instructions in the language used on the job, it would be inappropriate for a selection test to require proficiency in reading and writing that particular language.

Standard 9.9 When multiple versions of a test are intended to be comparable, test developers should report evidence of test comparability.

Comment: Evidence of test comparability may include but is not limited to evidence that the different language versions measure equivalent or similar constructs, and that score reliability and validity of inferences from scores from the two versions are comparable.

Standard 9.10 Inferences about test takers' general language proficiency should be based on tests that measure a range of language features, and not a single linguistic skill.

Comment: For example, a multiple-choice, pencil-and-paper test of vocabulary does not indicate how well a person understands the language when spoken nor how well the person speaks the language. However, the test score might be helpful in determining how well a person understands some aspects of the written language. In making educational placement decisions, a more complete range of communicative abilities (e.g., word knowledge, syntax) will typically need to be assessed.

Standard 9.11 **When an interpreter is used in testing, the interpreter should be fluent in both the language of the test and the examinee's native language, should have expertise in translating, and should have a basic understanding of the assessment process.**

Comment: Although individuals with limited proficiency in the language of the test should ideally be tested by professionally trained bilingual examiners, the use of an interpreter may be necessary in some situations. If an interpreter is required, the professional examiner is responsible for insuring that the interpreter has the appropriate qualifications, experience, and preparation to assist appropriately in the administration of the test. It is necessary for the interpreter to understand the importance of following standardized procedures, how testing is conducted typically, the importance of accurately conveying to the examiner an examinee's actual responses, and the role and responsibilities of the interpreter in testing.

By the Authority vested in me as President by the Constitution and the laws of the United States of America, and in order to advance the development of human potential, to strengthen the Nation's capacity to provide high-quality education, and to increase the opportunities for Hispanic Americans to participate in and benefit from Federal education programs, it is hereby ordered..."

Executive Order 12900
President Clinton, February 22, 1994

Recognizing the importance of increasing the level of educational attainment for Hispanic Americans, President Clinton established the White House Initiative on Educational Excellence for Hispanic Americans through Executive Order 12900 in September 1994. Guiding the White House Initiative is the President's Advisory Commission on Educational Excellence for Hispanic Americans, whose responsibility is to advise the president, the secretary of education and the nation on the most pressing educational needs of Hispanic Americans. The White House Initiative also provides the connection between the Commission, the White House, the federal government and the Latino community throughout the nation.

Current White House Initiative activities include initiating policy seminars, developing issue briefs, fact-sheets and information kits on the condition of Latinos in education, facilitating community outreach, increasing understanding and awareness of Hispanic-Serving Institutions (HSIs), and coordinating high-level efforts across the national government to improve education for Hispanics. These activities are driven by the president's request to assess:

- Hispanic educational attainment from pre-K through graduate and professional school;
- State, private sector, and community involvement in education;
- The extent to which federal activities in education complement existing efforts to increase education opportunities; and,
- Hispanic federal employment and federal recruitment strategies.

Accelerating the educational success of Hispanic Americans is among the most important keys to America's continued success. Please join us in ensuring educational excellence for all Americans.

The White House Initiative Staff

Sarita E. Brown
Executive Director

Deborah A. Santiago
Deputy Director

Richard Toscano
Special Assistant for Interagency Affairs

Deborah M. Montoya
Assistant to the Executive Director

Julie S. Laurel
Policy Analyst

Danielle Gonzales
Policy Intern